How Much of a Language Model Group's Collective Failure Is Visible in Its Pairs?

Anonymous authors
Paper under double-blind review

Abstract

How much do the pairwise failure rates of the models in a language-model group reveal about them failing together? Individual and pairwise rates, which form a pairwise table, do not in general identify the collective failure rate, the share of items on which every model fails. However, across 48 public evaluation pools from five releases, completions of the pairwise table approximate the collective failure rate closely. For groups of eight models, the range of collective failure rates allowed by the pairwise table is on average 16 percentage points wide, more than twice a group's observed rate, so the table alone pins that rate down only to within $\pm 120\%$ of its value. Yet the least-structured completion, pairwise maximum entropy, reconstructs that rate with 15.9% mean relative error (1.1 percentage points), and a Gaussian copula gives the same qualitative result. Where pools have enough items to measure it, most of the dependence among model failures is pairwise. The regularity has limits: error rises with group size and on held-out items when collective failures are scarce. One use follows. Across the 48 pools, groups chosen by maximum entropy from a pairwise table fail together less often on held-out items than groups of the most accurate models, on average and in most pools (3.3 percentage points at eight models), also when the choice is restricted to the strongest candidates. Exploratory studies on SWE-bench Verified and RewardBench 2 are consistent with this. The results hold within an evaluation, not across evaluations.

1 Introduction

Language models are increasingly used in groups. A coding pipeline may try several models and keep the first solution that passes its tests; a detector may flag an input when any of several models flags it. Such a group fails only when every member fails on the same item, so its reliability depends on whether the members make different mistakes, not only on how good each one is (Kleinberg & Raghavan, 2021; Bommasani et al., 2022; Toups et al., 2023). Leaderboards report individual scores and say nothing about shared mistakes, and independently built software has long been known to fail together more often than independence predicts (Eckhardt & Lee, 1985; Knight & Leveson, 1986).

A *pairwise table* adds one layer: each model's failure rate and, for every pair, how often both fail on the same item. Such tables are already used to measure correlated errors across providers, architectures, and capability levels (Kim et al., 2025; Goel et al., 2025). A group of three or more models is more than its pairs, however. We study its *collective failure rate*: the fraction of items on which every member fails. The pairwise table does not determine it, because two groups can have the same individual and pairwise rates and different collective failure rates. For the real group in Figure 1, any collective failure rate from 0 to 30% is consistent with its pairwise table.

Whether this ambiguity matters is therefore an empirical question. When real models are summarised by a pairwise table, how much of their collective failure is lost? If the loss is small, one table for M models describes every group that can be drawn from them, and redundancy can be studied without inspecting each group. If the loss is large, pairwise summaries can mislead. Public releases of full response matrices, which record every model's result on every item, let us measure the loss directly.

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

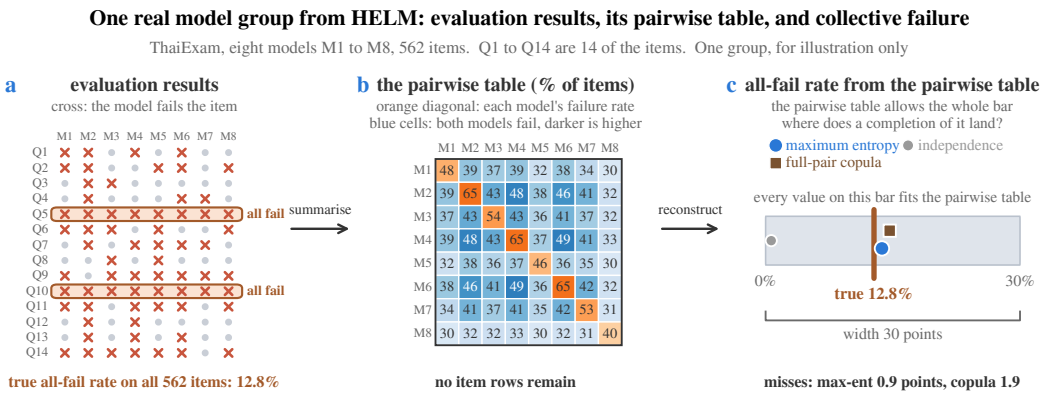


Figure 1: **One group, for illustration.** The collective failure rate is the share of items on which every model fails (**a**). The pairwise table keeps only each model's failure rate and each pair's co-failure rate (**b**). Any rate from 0 to 30% is consistent with it, so the table alone does not determine the collective failure rate (**c**). Two completions of the table, maximum entropy and a Gaussian copula, land 0.9 and 1.9 points from the true rate. The paper averages such misses over 13,877 groups (Section 3). This group was chosen by the display rule of Appendix B.

We study 48 model-by-benchmark evaluation pools from five public releases (Liang et al., 2023; Fourrier et al., 2024; Zhuang et al., 2025; White et al., 2025; Chen, 2026). Each pool contains 11 to 346 models evaluated on the same items. Our findings are:

- **The pairwise table does not pin down collective failure, yet real groups sit close to a simple completion of it.** At eight models, the exact range of collective failure rates allowed by the pairwise table (the compatible interval, Section 3) is wide, on average more than twice the observed rate, while maximum entropy reconstructs that rate with 15.9% mean relative error, or 1.1 percentage points, and a Gaussian copula gives the same qualitative result.
- **The regularity has limits.** Error grows with group size and when collective failure is rare. On held-out items at eight models, maximum-entropy error rises to 32.4%.
- **The pairwise table is useful for choosing a group.** Groups chosen from it by maximum entropy fail together less often on held-out items than groups of the individually most accurate models, on average and in most pools (3.3 percentage points at eight models), also under the strong-quarter rule, which chooses only among the top quarter of the candidates by accuracy and never fewer than twice the group size. Exploratory studies on SWE-bench Verified and RewardBench 2 are consistent with this (Section 5).

A supporting result concerns incomplete logs in which each model was scored on its own random subset of the items, so that a whole group shares few items while each pair shares many: maximum entropy then misses collective failure at five models by 24%, against 66% for direct counting on the shared items, although one common subset of the same size does better (14%, Section 3).

Our contribution is the measurement: the regularity, its limits, and one decision study, made with existing completions rather than a new estimator. The aligned response matrices of the 48 pools and the code are in the supplementary material.

2 Setup

For each item, let a model's failure indicator be one when it fails and zero when it succeeds. For a group of k models, the *collective failure rate*, denoted β, is the fraction of items on which all k indicators equal one. It is also called the all-fail rate. Its complement, *coverage*, is the fraction of items on which at least one member succeeds. A pairwise table contains each model's failure rate and each pair's co-failure rate, the fraction of items on which both fail. It does not contain the item-level indicators or β.

Table 1: **Pairwise-table completions compared with direct counting and a marginal-only baseline.** Values at eight models, sorted by the held-out column. Reconstruction uses all 48 pools and is given as mean relative error (%) and as mean absolute error in percentage points, each averaged within pools and then across pools. The held-out comparison uses the 39 pools of the balanced design (Appendix C); on all 48 pools the held-out error at eight models is 32.4% for maximum entropy and 30.3% for direct counting (Section 3).

		Reconstruction, 48 pools		Held out, 39 pools
Method	Input	(%)	(pp)	(%)
Direct empirical count	response matrix	0.0	0.00	16.1
Maximum entropy	pairwise table	15.9	1.12	18.8
Two-factor probit	pairwise table	23.2	1.43	19.8
Full-pair Gaussian copula	pairwise table	23.6	1.53	20.4
One-factor probit	pairwise table	26.4	1.76	22.9
Equicorrelated copula	pairwise table	30.9	1.98	22.9
Two-class latent model	pairwise table	39.7	3.10	31.9
Compatible-interval midpoint	pairwise table	60.6	2.05	33.2
Bahadur expansion	pairwise table	60.1	4.72	51.2
Independence	individual rates	88.2	10.89	83.4

An *evaluation pool* is a set of models scored on the same benchmark items. The 48 pools contain 108,572 item rows, counting an item once per pool (Appendix D notes repeated rows in EmbedLLM). For each pool and each size $k \in \{3, 4, 5, 6, 8\}$, we sample 60 groups uniformly from the available models. Groups with a model that always succeeds or always fails lie on a boundary, and relative error is undefined when no collective failure is observed. Removing those cases leaves 13,877 groups evaluated by every method (Appendix D reports their absolute-error sensitivity).

We separate two questions. *Reconstruction* builds the pairwise table and measures β on the same items. It asks what information the pairwise table preserves. *Held-out prediction* builds the pairwise table on one half of the items, the *fit half*, and measures β on the other, the *scored half*. It also includes sampling variation. Each split is used in both directions, so one split gives two *split directions*. Unless stated otherwise, averages are computed within each pool and then across pools. Groups from one pool are not treated as independent observations (Appendix E).

For each pairwise table, a linear program finds the smallest and largest possible values of β. We call them L and U. Every joint distribution matching the supplied individual and pairwise rates has $\beta \in [L, U]$ (Boole, 1854; Hailperin, 1965). We compare this identification result, which makes no distributional assumption, with completions that select one distribution compatible with, or fitted to, the same table. Appendix D defines all completions, the interval midpoint, and direct counting.

3 Collective failure is approximately recoverable from pairs

The compatible interval is wide. For each group, we measure the half-width $(U - L)/2$ relative to its observed collective failure rate. At eight models, this quantity averages 120%, and the interval itself averages 15.8 percentage points, from 6.9% to 22.6%. The interval need not be centred on the truth, and the half-width is a scale for how much the pairwise table leaves unresolved. Pairwise rates therefore do not determine collective failure.

At eight models the observed rate's median position in the interval, $(\beta - L)/(U - L)$, is 0.56 with an interquartile range of 0.40–0.67, and the interval midpoint still misses by 60.6% on average.

Observed collective failure rates nevertheless lie close to pairwise completions. Pairwise maximum entropy chooses the highest-entropy distribution matching the pairwise table (Jaynes, 1957). At eight models, its mean reconstruction error is 15.9%, or 1.1 percentage points. Its error is about one eighth of the half-width left by exact identification. Its median signed error is zero through five models and -1.8% of β at eight, where it lies below β on 59% of groups (Appendix D). A full-pair Gaussian copula, a different construction fitted to the same individual and pairwise rates, misses by 23.6%. Independence, which discards the co-failure rates, misses by 88.2%. Figure 2 shows the same pattern across all five releases.

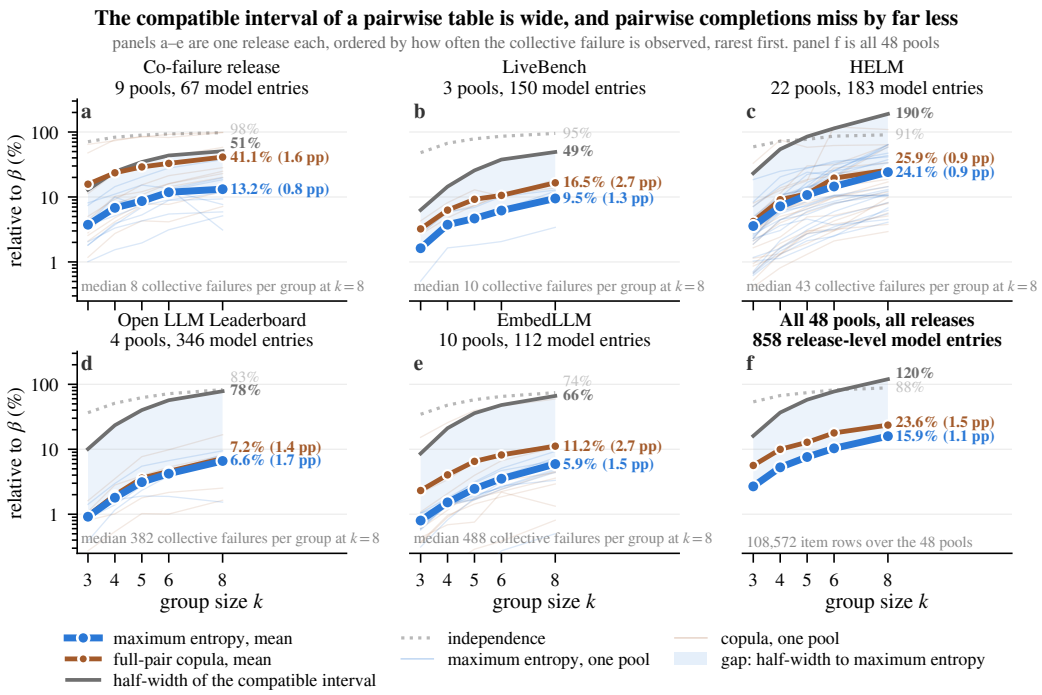


Figure 2: **The pairwise table permits a wide interval, while two pairwise completions lie much closer to the observed collective failure rate.** Grey is the mean half-width of the exact compatible interval, divided by the observed rate. Blue and brown are reconstruction errors on the same items. Thin lines are pools, and panel **f** averages all 48. Panel numbers: median collective failures per eight-model group in that release. Model entries: models listed by that release.

The result is not a proof that maximum entropy is the correct joint distribution. At eight models the copula is slightly more accurate when its fitted correlation matrix is valid (14.5% against 15.7% on the 70% of groups where it is), while maximum entropy is far more stable on the 30% of tables for which that matrix requires repair (15.4% against 43.6%, Appendix D.1). We therefore use maximum entropy as one concrete witness that the pairwise table can reconstruct the higher-order quantity, not as the contribution itself (Table 1).

Where the observation weakens. This result is strongest for groups of three to eight models. In a balanced design, which follows a fixed set of 39 pools and keeps only groups with at least 20 observed collective failures (10 from ten models upward) as the group size grows (Appendix C), maximum-entropy error rises from 1.9% at three models to 17.6% at fourteen (Figure 4a). Scale also matters: at eight models, maximum entropy misses by 61.6% when $\beta \leq 1\%$ and by 6.6% when $\beta > 20\%$, or 0.23 and 2.10 percentage points (Appendix D). Groups drawn from the strong-quarter candidates of each pool show the same errors as random groups (Appendix C).

Reconstruction should not be read as prediction on new items. Splitting every pool in half raises maximum-entropy error at eight models from 15.9% to 32.4%, and direct counting from the fit half gives 30.3% (Figure 3b). Table 1 reports the same comparison on the 39-pool cohort with more collective failures, where both errors are lower. Error falls by more than a factor of ten as the fit half contains more collective failures (Figure 3a). The main source of this degradation is that rare collective failures are measured noisily.

When the full matrix is incomplete. We scored each model on an independent random fraction f of a pool's items, as in pooled logs where each model was scored on its own random subset, so that a pair is measured on about f^2 of them and the whole group on about f^k. At $f = 0.5$ and five models, a median of 30 items are scored by every member against 236 by a pair. On the groups with such items, maximum entropy misses the all-item rate by 24.2% and direct counting on those items by 65.6%, and at eight models by 31.7% and 129.2%. This describes logs that are already incomplete,

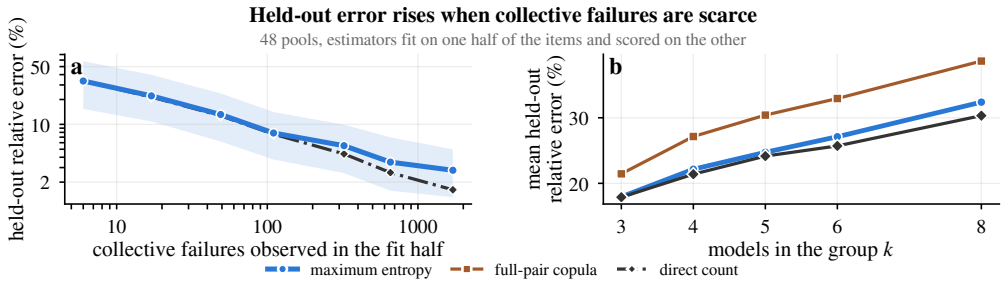


Figure 3: **Held-out error rises when collective failures are scarce. a** Median maximum-entropy error over groups, binned by the number of collective failures observed in the fit half, with the interquartile range shaded and direct counting for comparison. **b** Relative error averaged within each pool and then across all 48 pools. Direct counting uses the fit half's full response matrix.

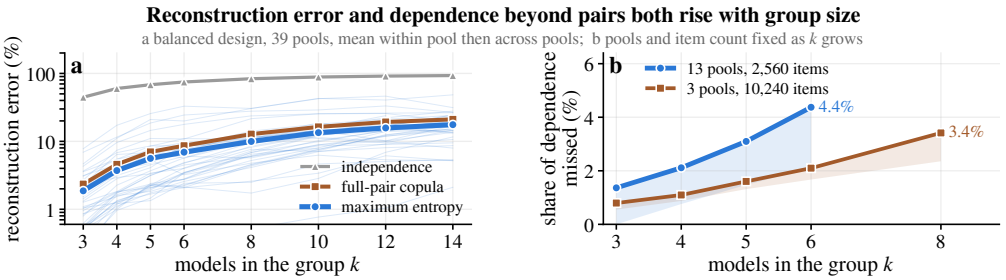


Figure 4: **Group size. a** Reconstruction error in the balanced design (Appendix C); thick lines are means over pools for maximum entropy, the full-pair copula, and independence, thin lines individual pools. **b** Share of a group's total dependence, measured as multi-information, left after matching every individual and pairwise rate; pools and item count are fixed as k grows. Lines use the plug-in entropy, and the shading extends to the Grassberger-corrected estimate (Appendix F).

not an evaluation design: one common subset with the same number of evaluations gives 13.6% and 17.3% by direct counting. Independently sampled tables are also often jointly infeasible, 29% at five models and 49% at eight, and the fitter then returns an approximate fit whose moments remain mismatched (Appendix I).

4 Dependence by interaction order

Accurate reconstruction of one tail event does not show that the whole failure distribution is pairwise. We measure this separately with an interaction-order decomposition (Schneidman et al., 2003). For a group of models, we first fit a maximum-entropy distribution matching only individual failure rates, then add every pairwise rate, then every third-order rate, and continue through the full empirical distribution. The reduction in entropy at each step measures the dependence first represented at that order.

This calculation requires enough items to estimate all possible failure patterns. We therefore use 13 pools with 2,560 items for groups of six models, and three pools with 10,240 items for groups of eight, with up to 60 groups per pool. Pairwise terms account for 95.6% of the measured dependence at six models, and 12 of the 13 pools exceed 94% (Figure 4b). The three eight-model pools, all MMLU-family (Open LLM Leaderboard MMLU-Pro, HELM MMLU, EmbedLLM MMLU), show the same pattern, with 96.6% at eight models. Standard finite-sample corrections increase rather than decrease the pairwise share.

The decomposition describes where the measured dependence sits in these groups. It is not a mechanism for why it sits there. By Pinsker's inequality it bounds the absolute error, $|\beta - \hat{\beta}_2| \leq \sqrt{(S_2 - S_k)/2}$ with entropies in nats (Appendix F), but the bound is loose, 13.8 points against an observed miss of 0.7 points at eight models, and gives no control of relative error on rare events.

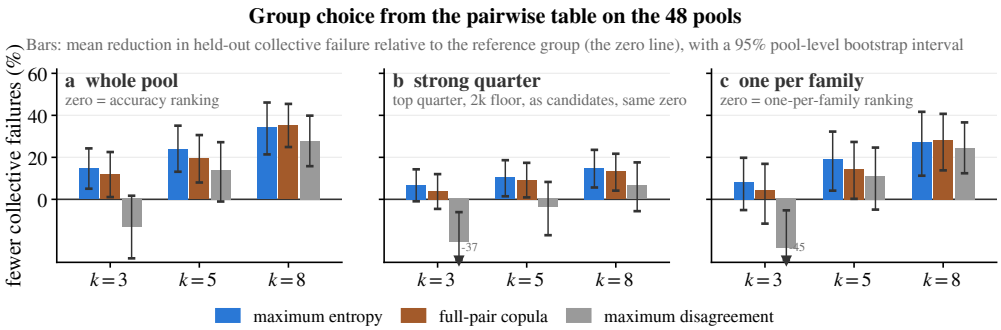


Figure 5: **Group choice from the pairwise table on the 48 pools.** Bars: mean reduction in held-out collective failure relative to the reference group, which is the zero line, with a 95% pool-level bootstrap interval. **a** Reference accuracy ranking. **b** Candidates under the strong-quarter rule (the top quarter of the candidates, with a $2k$ floor), same reference. **c** Reference one-per-family ranking, backfilled by accuracy when a pool has fewer than k families. Intervals that leave the panel are marked with their lower end.

5 Choosing groups for coverage

Suppose an evaluation reports a pairwise table and a user wants a group whose members cover one another's failures. Is selecting from that pairwise table better than taking the individually most accurate models?

Target and protocol. We measure the collective failure rate. This is the relevant event when an external check, such as unit tests, can recognise a successful answer, so one member suffices. It is not majority-vote error. Each pool is split in half. Individual and pairwise rates from the fit half select a group, whose realised collective failure is measured on the scored half. The halves are then reversed. We test groups of three, five, and eight models in the 48 pools. The reference group is *accuracy ranking*: the k models with the highest fit-half accuracy. Three selectors, each reading only the pairwise table, are compared with it (Appendix G adds the equicorrelated copula). Each selector builds its group one model at a time, adding the candidate that most lowers its estimated collective failure rate or, for maximum pairwise disagreement (maximum disagreement below, Kuncheva & Whitaker, 2003), most raises the mean pairwise disagreement. Results are the change in held-out collective failure relative to the reference, averaged within each pool and then across pools, with a pool-level bootstrap interval (Appendix G gives every selector, setting, and pool).

Results on the 48 pools. Maximum-entropy selection reduces held-out collective failure relative to accuracy ranking by 15%, 24%, and 34% at three, five, and eight models, or 2.8, 3.4, and 3.3 percentage points, and the interval excludes zero at every size (Figure 5a, Appendix G). The full-pair copula gives similar reductions. The two completions are close, so the experiment supports the information in the pairwise table rather than a property of one completion. Maximum disagreement, which reads the same pairs through a heuristic, is worse than accuracy ranking at three models and below both completions at eight. Resampling the five releases rather than pools gives wider intervals and lower means, and only the eight-model reduction is clearly separated from zero (Appendix G).

Restriction to strong models. Unrestricted selection sometimes adds a weak model, because even chance success can improve coverage: on HELM MMLU at five models, maximum entropy adds two OLMo-7B models, lowering collective failure while the members' mean accuracy falls from 86% to 64% (Appendix G). We therefore repeated the experiment with every candidate set restricted by the *strong-quarter rule* to the top quarter of its M candidates with a floor of $2k$, that is, the $\max(\lceil M/4 \rceil, 2k)$ most accurate, with the rule fixed before the comparison was run. Accuracy ranking is unchanged, because its members already lie in that set. The reductions become 7%, 10%, and 15%, and the interval excludes zero at five and eight models (Figure 5b, Table 10).

A one-per-family reference. Groups are also often spread across model families. We label each model by the architecture class of its weights or, for API models, by the lab that pretrained its base,

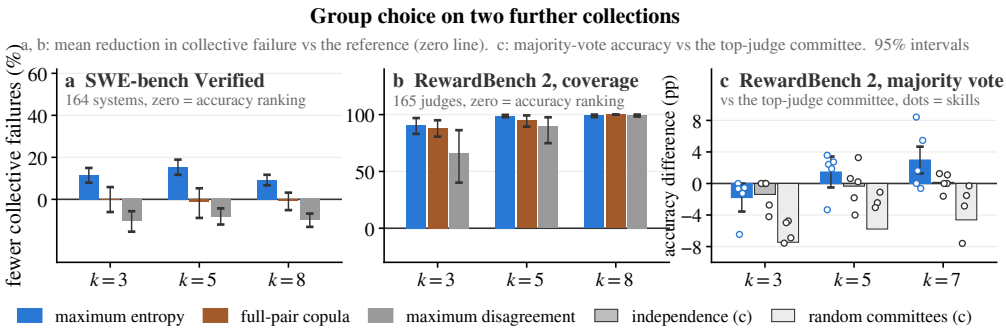


Figure 6: **Group choice on two further collections.** **a** SWE-bench Verified, 164 systems over 20 repeated splits of the same tasks, reference accuracy ranking, interval over splits. **b** RewardBench 2, 165 judges over five skills, reference accuracy ranking, interval over split directions; the event medians quoted in the text are not the quantities behind these bars. **c** RewardBench 2 committees chosen by maximum entropy, under independence, and at random, evaluated by majority vote on a held-out skill, in points of strict prompt accuracy against the top-judge committee, the majority vote of the k judges with the highest accuracy on the source skills, with a prompt-cluster bootstrap interval for maximum entropy. Dots are the five skills. In **a** and **b**, bars are mean reductions and positive values mean fewer collective failures. Intervals that leave the panel are marked with their lower end.

and build a *one-per-family ranking* that walks down the accuracy ranking taking at most one model per family, backfilling by accuracy when the pool has fewer than k families. It lowers collective failure by 3 to 4% relative to accuracy ranking, with intervals including zero, and maximum-entropy selection is lower than it at every size while using fewer families (Figure 5c, Table 10).

Two further collections. We applied the same steps to two collections outside the 48 pools (Appendix G). SWE-bench Verified (Jimenez et al., 2024; OpenAI, 2024) has 164 systems on 500 repository tasks, and a system succeeds only if its patch passes the repository's tests (Appendix H). Over twenty repeated random splits of the same 500 tasks, maximum-entropy groups fail together on 9 to 15% fewer held-out tasks than the most accurate systems, with similar values when near-duplicate submissions are not merged, and are lower on most of the 40 split directions (Table 12, Appendix H). The interval reflects variability over re-splits of these 500 tasks, not uncertainty over new tasks. On the 41 systems that share one agent, the reductions are smaller and the eight-model gain does not survive (Appendix H). The full-pair copula gives no reduction on this collection, and maximum disagreement is worse than accuracy ranking (Figure 6a). RewardBench 2 (Malik et al., 2026) scores 165 reward models on preference comparisons in five skills. A panel of judges fails together on a comparison when every judge scores the rejected response at least as high as the chosen one. Panels of the most accurate judges fail together on a median of 30, 22, and 16 held-out comparisons at three, five, and eight judges; panels chosen by maximum entropy or by the full-pair copula on a median of 1.5, 0, and 0 (Figure 6b, Table 12). These reductions are descriptive, since the collection has five skills and few collective failures.

The same collection allows an exploratory test on a decision that needs no external check. Committees are chosen from the pairwise table of four skills, by the lowest estimated rate of a wrong majority, and evaluated by majority vote on the fifth. This is a committee evaluation derived from the released judge scores, not a RewardBench leaderboard result. The measure is strict prompt accuracy, the share of prompts on which the majority ranks the chosen response above each of its distractors. Committees chosen by maximum entropy are behind the top-judge committee, the majority vote of the k judges most accurate on the source skills, by 1.8 points at three judges and ahead at five and seven, where the gain is 3.0 points with a prompt-cluster bootstrap interval of $[1.3, 4.7]$ (Figure 6c). Committees chosen under independence and random committees score lower (Table 12).

Scope of the application. Selection depends on ordering candidate groups, not on estimating every collective failure rate exactly. The held-out experiment evaluates the realised choice, so a calibration error that is consistent across groups does not by itself change the decision. When the complete response matrix is available, counting collective failure directly from it is marginally better than maximum entropy, as expected, by 0.08 to 0.15 percentage points in a controlled comparison

in which every selector scores the same candidate groups (Appendix G). Pairwise selection is therefore not a replacement for direct counting: the comparison measures how much selection signal survives compression to the pairwise table. Every reconstruction and coverage result is within the evaluation that supplied the pairwise table, and the majority-vote study is the one leave-one-skill-out experiment (Appendix I).

6 Related work

Bounds and completions. Bounding a joint probability from low-order statistics goes back to Boole and Fréchet and is made sharp by Hailperin's linear program (Boole, 1854; Fréchet, 1951; Hailperin, 1965). Pairwise maximum entropy is the Ising model of statistical physics: Schneidman et al. (2006) found that it captures most of the collective activity of retinal neurons, later work mapped its limits as populations grow (Roudi et al., 2009; Tkačik et al., 2014), and Ising models have been used to aggregate the labels of several language-model judges (Balasubramanian et al., 2026). Appendix A reviews this lineage and the other completions of Table 1.

Correlated failure of language models. Common-mode failure is classical in reliability: independently developed software versions fail together more often than independence predicts (Eckhardt & Lee, 1985; Knight & Leveson, 1986; Littlewood & Miller, 1989), and common-cause models exist for redundant components (Atwood, 1986). For language models, correlated errors across providers, architectures, and capability levels have been measured from pairwise agreement (Kim et al., 2025; Goel et al., 2025), outputs have been found to be homogeneous across models (Jiang et al., 2025), and judges favour models related to themselves (Panickssery et al., 2024; Li et al., 2026; Awuni et al., 2026). Chen (2026) proves that pairwise correlations cannot identify β for three or more models, the non-identification measured here, and finds on 67 frontier models that a full pairwise Gaussian copula, projected to the nearest positive semidefinite matrix, predicts $\beta = 0.023$ against 0.052 observed on open-ended MATH-500, with an underpricing ratio near one at two models that rises with pool size and is attributed to a common-mode component. Re-asking GPQA-Diamond in free-response form opens an all-wrong tail ($\beta = 0.127$) absent in multiple choice. Our groups have three to eight models, where that ratio is near one, our error also rises with group size (Appendix C), the repaired-copula gap of Appendix D.1 is consistent with that projection, and on the release's MATH-500, MATH-hard, CodeGen, and GPQA-open pools maximum entropy misses β at eight models by 5.9%, 11.8%, 9.3%, and 5.1% (Table 3), with a median signed error of -1.8% of β across all pools (Appendix D). Relative to these studies, this paper measures the exact compatible interval and the reconstruction error side by side across many public pools, quantifies where the approximation weakens with group size, event rarity, and held-out items, and tests whether the information left in the pairwise summary is enough for group choice.

Combining models. Ensembles gain from members that err differently (Krogh & Vedelsby, 1994; Breiman, 1996; Dietterich, 2000). Classical diversity measures include pairwise ones, such as the disagreement rate that Section 5 uses as a selector, and group-level ones, and diversity alone is an unreliable predictor of ensemble accuracy (Kuncheva & Whitaker, 2003). For majority voting, second-order PAC-Bayes bounds use the tandem loss (Masegosa et al., 2020) and the C-bound the first two moments of the voting margin (Germain et al., 2015). Label aggregation from several noisy sources models their error rates and, in data programming, their pairwise dependencies (Dawid & Skene, 1979; Ratner et al., 2016; Bach et al., 2017). For language models, cascades, routers, response fusion, mixtures of agents, debate, and judge panels exploit differences in where models succeed (Chen et al., 2024; Jiang et al., 2023; Ong et al., 2025; Wang et al., 2025; Du et al., 2024; Verga et al., 2024; Yun et al., 2026), a literature surveyed by Chen et al. (2026), and routers and judges are themselves benchmarked (Lu et al., 2026; Zheng et al., 2023). Which models to combine has been studied with diversity objectives (Tekin et al., 2024; Ali, 2026), with per-item feedback (Neoh et al., 2026; Zhang et al., 2026), with cross-model consensus (Liu, 2026), and with the mutual information between labels and predictions under a query budget (Turkmen et al., 2026). Section 5 differs in its inputs and its target. The selectors see only the pairwise table, and the target is the collective failure rate of a fixed group of different models, which is the coverage endpoint. Because selection on estimated rates favours optimistic estimates (Smith & Winkler, 2006), every chosen group is scored on unseen items.

Evaluation summaries. Item-level results released by HELM, the Open LLM Leaderboard, EmbedLLM, LiveBench, and the co-failure release (Liang et al., 2023; Fourrier et al., 2024; Zhuang et al., 2025; White et al., 2025; Chen, 2026) make this study possible. Item-response theory models the response matrix through latent ability and item parameters (Rasch, 1960; Lord & Novick, 1968), and anchor-point and sparse-benchmark methods estimate model performance from fewer items (Vivek et al., 2024; Maia Polo et al., 2024; Kipnis et al., 2025). The pairwise table instead keeps every item but discards item identity, the loss Section 3 measures.

7 Discussion

Across the evaluations studied here, a pairwise table of individual and pairwise failure rates preserves much of the information needed to reconstruct one higher-order event. This does not follow from probability theory, since the exact compatible interval remains wide. It is a regularity of the observed model populations, and it means that pairwise summaries can be used to study redundancy without inspecting every model group.

Because a pairwise table contains no per-question records, it could also serve as an intermediate release in settings where item-level test data are withheld (Jacovi et al., 2023; Zhao et al., 2025), a use we have not evaluated. Aggregate releases can still permit inference or reconstruction in other settings (Dobra & Fienberg, 2000; Narayanan & Shmatikov, 2008).

The finding is dataset-dependent: a pairwise table measured on one benchmark does not predict collective failure on an unrelated benchmark (Appendix I), so the relevant interpretation is within an evaluation population.

8 Conclusion

Individual and pairwise failure rates do not determine how often an entire model group fails, and the exact compatible interval stays wide. Across 48 public evaluation pools, however, the observed collective failure rate lies much closer to pairwise completions than that interval would suggest, and in data-rich pools pairwise interactions account for most of the measured dependence.

The regularity is approximate. It weakens for larger groups, rare collective failures, and held-out items with limited sampling support. Within an evaluation, groups chosen by maximum entropy from a pairwise table fail together less often on held-out items than groups ranked by accuracy, on average and in most pools (3.3 percentage points at eight models), also under the strong-quarter rule. Exploratory studies of two further collections are consistent with this.

AI use statement

In this work we used generative AI tools for literature search, for building the aligned response matrices from the public releases, for producing the figures and tables, for feedback on the experimental design and on the interpretation of results, and for editing the text. We checked the resulting analyses, references, and figures. Every number in the paper is produced by a script in the supplementary material and was regenerated from the raw response matrices for this version, and every reference was checked against the publisher's record. No data were generated by AI; all data are public evaluation records. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

This work uses only public evaluation records of language models, released by their authors under their own terms, with no recorded outcome altered (the deduplication rules are in Appendices B and H). It involves no human subjects, no personal data, and no new data collection. The models studied are identified by the names their releases give them, and we make no claim about any individual model beyond what its public evaluation record shows.

The quantity we study has a dual character we think worth stating. A method for choosing groups of models that fail together less often can be used to build more reliable systems, and the same method

identifies which models are redundant with each other, which could be used to justify narrowing a model ecosystem rather than broadening it. Our results speak only to collective failure on the benchmark that supplied the statistics, and Appendix I reports a pre-specified null showing they do not transfer across benchmarks. They should not be read as evidence that any particular group of models is safe to deploy.

Reproducibility statement

Every pool comes from a public release, listed with its source in Appendix B, with no recorded outcome altered. The two duplicate or near-duplicate matrices excluded in Appendix B are named there with the reason. How groups are drawn, and the criteria that define the balanced design, are given in Section 2 and Appendix C, with the nine excluded pools and their qualification failures listed in full. Every completion of Table 1, how each is fitted, and the repair applied to the copula are in Appendix D. The unit of analysis, the bootstrap, and the pairing used for method comparisons are in Appendix E, together with what treating groups rather than pools as independent would have done to the intervals. The supplementary material contains the 48 aligned response matrices as derived binary data, one row per item and one column per model, with the model identifiers of each pool and the provenance of each release, together with the code that builds them from the public releases, computes every estimator and every bound, and regenerates each figure and table in this paper. The code is released under the MIT licence and the derived files under CC BY 4.0, subject to the terms of the source releases.

References

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

Junade Ali. Quantifying diversity of thought: A predictive law of weighted LLM ensemble lift. *arXiv preprint arXiv:2607.17384*, 2026.

Corwin L. Atwood. The binomial failure rate common cause model. *Technometrics*, 28(2):139–148, 1986.

David Ababio Awuni, Luke E. K. Achenie, Benjamin Tei Partey, Elvis Gyasi Owusu, and Nii-Nai Derrick Sowah. Who judges matters: Measuring family-conditioned preference in LLM-as-judge panels. *arXiv preprint arXiv:2609.17857*, 2026.

Stephen H. Bach, Bryan He, Alexander Ratner, and Christopher Ré. Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 273–282, 2017.

R. R. Bahadur. A representation of the joint distribution of responses to n dichotomous items. In Herbert Solomon (ed.), *Studies in Item Analysis and Prediction*, pp. 158–168. Stanford University Press, 1961.

Krishnakumar Balasubramanian, Aleksandr Podkopaev, and Shiva Prasad Kasiviswanathan. Dependence-aware label aggregation for LLM-as-a-judge via Ising models. *arXiv preprint arXiv:2601.22336*, 2026.

R. Darrell Bock, Robert Gibbons, and Eiji Muraki. Full-information item factor analysis. *Applied Psychological Measurement*, 12(3):261–280, 1988.

Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems 35*, pp. 3663–3678, 2022.

George Boole. *An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities*. Walton and Maberly, London, 1854.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Josef Chen. When does combining language models help? A co-failure ceiling on routing, voting, and mixture-of-agents across 67 frontier models. *arXiv preprint arXiv:2606.27288*, 2026.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024.

Zhijun Chen, Xiaodong Lu, Jingzheng Li, Pengpeng Chen, Zhuoran Li, Kai Sun, Yuankai Luo, Qianren Mao, Ming Li, Likang Xiao, Dingqi Yang, Yikun Ban, and Hailong Sun. Harnessing multiple large language models: A survey on LLM ensemble. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 7807–7816, 2026.

Anders Christoffersson. Factor analysis of dichotomized variables. *Psychometrika*, 40(1):5–32, 1975.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Xiang Deng, Jeff Da, Edwin Pan, Yannis Yiming He, Charles Ide, Kanak Garg, Niklas Lauffer, Andrew Park, Nitin Pasari, Chetan Rane, Karmini Sampath, Maya Krishnan, Srivatsa Kundurthy, Sean Hendryx, Zifan Wang, Vijay Bharadwaj, Jeff Holm, Raja Aluri, Chen Bo Calvin Zhang, Noah Jacobson, Bing Liu, and Brad Kenstler. SWE-Bench Pro: Can AI agents solve long-horizon software engineering tasks? *arXiv preprint arXiv:2509.16941*, 2025.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, Lecture Notes in Computer Science, pp. 1–15. Springer, 2000.

Adrian Dobra and Stephen E. Fienberg. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences*, 97(22): 11885–11892, 2000. doi: 10.1073/pnas.97.22.11885.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763, 2024.

Dave E. Eckhardt and Larry D. Lee. A theoretical basis for the analysis of multiversion software subject to coincident errors. *IEEE Transactions on Software Engineering*, SE-11(12):1511–1517, 1985.

Lawrence J. Emrich and Marion R. Piedmonte. A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45(4):302–304, 1991.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open LLM Leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`, 2024.

Maurice Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon, 3e série, Sciences, Section A*, 14:53–77, 1951.

Alan Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2):141–149, 1992.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16:787–860, 2015.

Shashwat Goel, Joschka Strüber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines AI oversight. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19621–19678, 2025.

Michael B. Gordy. A comparative anatomy of credit risk models. *Journal of Banking & Finance*, 24(1–2):119–149, 2000.

Peter Grassberger. Entropy estimates from insufficient samplings. *arXiv preprint physics/0307138*, 2003.

Theodore Hailperin. Best possible inequalities for the probability of a logical function of events. *The American Mathematical Monthly*, 72(4):343–359, 1965.

Nicholas J. Higham. Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, 22(3):329–343, 2002. doi: 10.1093/imanum/22.3.329.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5075–5084, 2023.

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, 2023.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems 38, Datasets and Benchmarks Track*, 2025.

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

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, 2025.

Alex Kipnis, Konstantinos Voudouris, Luca M. Schulze Buschoff, and Eric Schulz. metabench – a sparse benchmark of reasoning and knowledge in large language models. In *International Conference on Learning Representations*, 2025.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

John C. Knight and Nancy G. Leveson. An experimental evaluation of the assumption of independence in multiversion programming. *IEEE Transactions on Software Engineering*, SE-12(1): 96–109, 1986.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems 7*, pp. 231–238. MIT Press, 1994.

Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

Seymour M. Kwerel. Bounds on the probability of the union and intersection of m events. *Advances in Applied Probability*, 7(2):431–448, 1975.

Paul F. Lazarsfeld and Neil W. Henry. *Latent Structure Analysis*. Houghton Mifflin, Boston, 1968.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference leakage: A contamination problem in LLM-as-a-judge. In *International Conference on Learning Representations*, 2026.

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

Bev Littlewood and Douglas R. Miller. Conceptual modeling of coincident failures in multiversion software. *IEEE Transactions on Software Engineering*, 15(12):1596–1614, 1989.

Ning Liu. LLMs as a jury: Cross-model consensus can outperform process reward models for LLM reasoning. *arXiv preprint arXiv:2607.10139*, 2026.

Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, 1968.

Yifan Lu, Rixin Liu, Jiayi Yuan, Xingqi Cui, Shenrun Zhang, Hongyi Liu, and Jiarong Xing. RouterArena: An open platform for comprehensive comparison of LLM routers. In *International Conference on Learning Representations*, 2026.

Jakob H. Macke, Manfred Opper, and Matthias Bethge. Common input explains higher-order correlations and entropy in a simple model of neural population activity. *Physical Review Letters*, 106 (20):208102, 2011.

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, 2024.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing reward model evaluation. In *International Conference on Learning Representations*, pp. 144839–144866, 2026.

Andrés R. Masegosa, Stephan S. Lorenzen, Christian Igel, and Yevgeny Seldin. Second order PAC-Bayesian bounds for the weighted majority vote. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5263–5273, 2020.

George A. Miller. Note on the bias of information estimates. In Henry Quastler (ed.), *Information Theory in Psychology: Problems and Methods*, pp. 95–100. Free Press, Glencoe, IL, 1955.

Jordi Molins and Eduard Vives. Model risk on credit risk. *Risk and Decision Analysis*, 6(1):65–78, 2016.

Bengt Muthén. Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43(4): 551–560, 1978.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy*, pp. 111–125, 2008. doi: 10.1109/SP.2008.33.

Tzeh Yuan Neoh, Nicholas Teh, Je Qin Chooi, Paul W. Goldberg, and Milind Tambe. Efficient ensemble selection from binary and pairwise feedback. *arXiv preprint arXiv:2605.09588*, 2026.

Nils J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28(1):71–87, 1986.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations*, pp. 34433–34448, 2025.

OpenAI. Introducing SWE-bench Verified. `https://openai.com/index/introducing-swe-bench-verified/`, 2024.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems 37*, pp. 68772–68802, 2024.

András Prékopa. Boole-Bonferroni inequalities and linear programming. *Operations Research*, 36 (1):145–162, 1988.

Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen, 1960.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems 29*, 2016.

Yasser Roudi, Sheila Nirenberg, and Peter E. Latham. Pairwise maximum entropy models for studying large biological systems: When they can work and when they can't. *PLoS Computational Biology*, 5(5):e1000380, 2009.

Elad Schneidman, Susanne Still, Michael J. Berry, and William Bialek. Network information and connected correlations. *Physical Review Letters*, 91(23):238701, 2003.

Elad Schneidman, Michael J. Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440:1007–1012, 2006.

Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

James E. Smith and Robert L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

Selim Furkan Tekin, Fatih Ilhan, Tiansheng Huang, Sihao Hu, and Ling Liu. LLM-TOPLA: Efficient LLM ensemble by maximising diversity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11951–11966, 2024.

Gašper Tkačik, Olivier Marre, Dario Amodei, Elad Schneidman, William Bialek, and Michael J. Berry II. Searching for collective behavior in a large network of sensory neurons. *PLoS Computational Biology*, 10(1):e1003408, 2014.

Connor Toups, Rishi Bommasani, Kathleen Creel, Sarah Bana, Dan Jurafsky, and Percy Liang. Ecosystem-level analysis of deployed machine learning reveals homogeneous outcomes. In *Advances in Neural Information Processing Systems 36*, pp. 51178–51201, 2023.

Yigit Turkmen, Baturalp Buyukates, and Melih Bastopcu. Don't always pick the highest-performing model: An information theoretic view of LLM ensemble selection. *arXiv preprint arXiv:2602.08003*, 2026.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1576–1601, 2024.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *International Conference on Learning Representations*, pp. 33944–33963, 2025.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations*, pp. 91595–91631, 2025.

Heecheol Yun, Kwangmin Ki, Jung Hyun Lee, and Eunho Yang. When to ensemble: Identifying token-level points for stable and fast LLM ensembling. In *International Conference on Learning Representations*, 2026.

Yichi Zhang, Kevin Lu, Yuang Zhang, Jie Gao, Lirong Xia, and Fang-Yi Yu. Mixture of complementary agents for robust LLM ensemble. *arXiv preprint arXiv:2605.24048*, 2026.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, and Furu Wei. MMLU-CF: A contamination-free multi-task language understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13371–13391, 2025.

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

Chunsheng Zhou. An analysis of default correlations and multiple defaults. *The Review of Financial Studies*, 14(2):555–576, 2001.

Richard Zhuang, Tianhao Wu, Zhaojin Wen, Andrew Li, Jiantao Jiao, and Kannan Ramchandran. EmbedLLM: Learning compact representations of large language models. In *International Conference on Learning Representations*, pp. 76913–76926, 2025.

A Related work on bounds and completions

Bounds from low-order statistics. Bounding the probability of a compound event from partial probabilistic information goes back to Boole (Boole, 1854), and Fréchet derived bounds on joint distributions with given margins (Fréchet, 1951). Hailperin showed that the sharpest bounds under specified event probabilities are the optimum of a linear program (Hailperin, 1965), an approach later developed as probabilistic logic (Nilsson, 1986) and as Boole–Bonferroni bounds from aggregated low-order intersections (Kwerel, 1975; Prékopa, 1988). The interval $[L, U]$ of Section 2 is Hailperin's program applied to the collective failure event.

Pairwise maximum entropy and interaction order. Maximum entropy under moment constraints gives an exponential-family distribution (Jaynes, 1957); for binary variables with individual and pairwise moments it is the Ising model, or a fully visible Boltzmann machine (Ackley et al., 1985; Wainwright & Jordan, 2008). Schneidman et al. (2006) found that such models capture most of the collective activity of retinal neurons. Later work analysed their limits as the population grows and fitted them to larger populations (Roudi et al., 2009; Tkačik et al., 2014), and showed that common input alone can produce higher-order correlations (Macke et al., 2011). The interaction-order decomposition of Section 4 is the connected-information measure of Schneidman et al. (2003). Ising models have also been used to aggregate the labels of several language-model judges (Balasubramanian et al., 2026).

Other completions of a pairwise table. Several model families define a joint distribution of correlated binary variables, although their usual estimators read different inputs. Copulas separate margins from dependence (Sklar, 1959), thresholded multivariate normals give binary variables with specified margins and correlations (Emrich & Piedmonte, 1991), and a Gaussian copula has been fitted to language-model predictions (Turkmen et al., 2026). Limited-information factor analysis of binary items uses first- and second-order margins (Christoffersson, 1975; Muthén, 1978), whereas full-information item factor analysis fits complete response patterns (Bock et al., 1988). Latent class models (Lazarsfeld & Henry, 1968), the Bahadur representation, exact through all orders and here

truncated after the pairwise terms (Bahadur, 1961), and models of joint default in credit risk (Gordy, 2000; Zhou, 2001; Molins & Vives, 2016) are further representations of multivariate binary dependence. Table 1 fits pairwise-table versions of these families as baselines; the gap between maximum entropy and the full-pair copula arises mostly on groups whose copula correlation matrix requires repair (Appendix D.1).

B POOLS

Table 2 lists every pool. In that table, n is the number of items, M the number of models, "Err" the mean model failure rate, β_8 the median collective failure rate for eight-model groups, and "Design" whether the pool enters the balanced analysis. Two further matrices of the co-failure release were removed as duplicates and are not counted anywhere in the paper: its MATH-500 run `E3`, which repeats the 200 items and 8 models of the `E2` MATH-500 run exactly, and its MATH-500 run `A2`, which is a re-run of 120 of those items on 15 of those models and agrees on 92% of cells. In the two LiveBench code pools, five models listed under two spellings differing only in letter case are counted once.

Table 3 gives the single-pool curves of Figure 2 and Table 4 the release curves drawn over them. Between them they hold every value the figure draws, so the figure can be rebuilt from the tables. The single-pool table reports relative error averaged over groups. The release table first averages within each pool and then across pools. Its "entries" count is the number of distinct models within a release; models appearing in two releases count once in each.

The group of Figure 1. Among the eight-model groups of Section 2 with an observed collective failure rate between 10% and 40%, a compatible interval at least as wide, relative to the observed rate, as the median group, and a maximum-entropy relative error at least two points below the full-pair copula's, the figure shows the group whose maximum-entropy error is nearest the median over all eight-model groups. Its models are the HELM entries claude-3-haiku-20240307, command-r-plus, seallm-7b-v2.5, gpt-3.5-turbo-0125, gpt-4o-mini-2024-07-18, openthaigpt-1.0.0-13b-chat, sailor-14b-chat, and llama-3-typhoon-v1.5x-70b-instruct, in the order M1 to M8. The fourteen items are a random sample that contains two collective failures, the nearest integer to the group's 12.8% rate on fourteen items. The compatible interval is the linear program of Section 2 on that group's pairwise table.

C THE BALANCED DESIGN

Figure 4a gives reconstruction error against group size. Maximum entropy runs from 1.87% at three models to 17.60% at fourteen, the full-pair copula from 2.35% to 21.07%, and independence from 44.6% to 93.0%. In percentage points maximum entropy runs from 0.31 to 2.15 and the copula from 0.55 to 1.96, so the ratio between them, taken from unrounded values, falls from 1.81 to 0.91 in points and from 1.26 to 1.20 in relative error. The mean β over the same range falls from 32% to 14%.

A group passes if no member has a failure rate of exactly 0 or 1 and the group fails together on at least 20 items. Qualification is decided at the five sizes $k \in \{3, 4, 5, 6, 8\}$: at each one, 400 candidate groups are drawn with a fixed seed, and a pool enters only if at least 30 pass at all five sizes. The first 30 passing groups form that pool–size combination. Thirty-nine of the 48 pools qualify. Table 5 gives the counts for the nine that do not.

The design is then extended to $k \in \{10, 12, 14\}$ on the qualifying pools, with the threshold lowered to 10 collective failures because the event is rarer at those sizes. A pool–size combination keeps up to 30 passing groups and is dropped altogether if fewer than 5 pass. That extension is why the total is 9,163 rather than $39 \times 8 \times 30 = 9{,}360$. All 39 pools fill every group size through $k = 8$. The 197 missing groups are all at $k \geq 10$ and come from five pools: GPQA-open has only 11 models, so k=12 and k=14 are impossible and k=10 admits just 11 distinct groups (79 missing groups). GPQA (E2) at k=14, Dyck at k=12 and LegalSupport at k=12 yielded 15, 8 and 9 passing candidates out of 400 (58 missing). Dyck and LegalSupport at k=14 yielded fewer than 5, so both cells were dropped (60 missing).

Table 2: **The 48 evaluation pools used in the main analysis.** n: items. M: models. Err: mean model failure rate. β_8: median collective failure rate of eight-model groups. Design: enters the balanced design. Releases: HELM (Liang et al., 2023), the Open LLM Leaderboard (Fourrier et al., 2024), EmbedLLM (Zhuang et al., 2025), LiveBench (White et al., 2025), and the co-failure release (Chen, 2026), whose run identifiers A2, E2, CG, and MH are kept. LCB is LiveCodeBench.

Release	Pool	n	M	Err	β_8	Design
Co-failure release	ARC (A2)	120	15	6%	0.8%	–
	CodeGen (CG)	63	18	56%	12.7%	–
	GPQA (E2)	130	52	59%	8.1%	✓
	GPQA-open	130	11	61%	27.7%	✓
	GSM8K (A2)	120	15	7%	0.8%	–
	MATH-500 (E2)	200	52	30%	10.0%	✓
	MATH-hard (MH)	298	67	26%	6.5%	✓
	MMLU (A2)	120	15	24%	3.3%	–
	MMLU-Pro (E2)	200	52	35%	4.0%	–
LiveBench	Coding completion	50	149	64%	18.0%	–
	LCB generation	78	150	59%	18.6%	✓
	Typos	100	58	47%	10.0%	–
HELM	AIR-Bench	5,000	41	38%	5.9%	✓
	BBQ	999	42	59%	1.8%	✓
	BoolQ	3,206	67	28%	1.2%	✓
	CivilComments	5,000	67	46%	0.5%	✓
	Commonsense	498	91	17%	0.6%	–
	Data imputation	424	67	27%	7.5%	✓
	Dyck	500	69	49%	1.6%	✓
	Entity matching	1,396	67	26%	0.1%	–
	GSM8K	1,000	90	33%	1.7%	✓
	IMDB	4,402	27	10%	0.9%	✓
	LSAT	454	69	80%	19.6%	✓
	LegalBench	2,047	89	46%	9.9%	✓
	LegalSupport	594	69	48%	0.7%	✓
	MATH	437	91	41%	3.0%	✓
	MMLU	14,042	69	28%	4.3%	✓
	MedQA	1,000	91	39%	4.7%	✓
	RAFT	1,348	67	42%	6.4%	✓
	Synthetic reasoning	3,000	69	77%	35.9%	✓
	ThaiExam	562	40	54%	10.1%	✓
	TruthfulQA	1,895	67	74%	17.4%	✓
	WikiFact	5,000	66	77%	50.0%	✓
	bAbI	3,572	70	69%	23.3%	✓
Open LLM Leaderboard	GPQA	546	346	66%	16.0%	✓
	MATH-hard	1,324	304	80%	43.2%	✓
	MMLU-Pro	11,994	339	54%	20.6%	✓
	MuSR	756	346	56%	25.6%	✓
EmbedLLM	ASDiv	2,304	112	96%	82.2%	✓
	GPQA	5,960	112	89%	66.4%	✓
	GSM8K	1,319	112	61%	11.0%	✓
	LogiQA	651	112	71%	19.2%	✓
	MMLU	14,015	112	47%	7.3%	✓
	MathQA	2,985	112	67%	29.2%	✓
	MedMCQA	4,183	112	58%	9.0%	✓
	PIQA	1,803	112	22%	8.7%	✓
	SocialIQA	1,930	112	67%	31.1%	✓
	TruthfulQA	817	112	64%	34.2%	✓

The exclusion is not neutral. The retained pools have a median of 1,324 items and a median model error rate of 56.3%. The excluded pools have 120 items and 26.5%. The balanced design therefore describes pools in which collective failure is comparatively common, and any number reported on it should be read that way. Because the design selects groups on the number of collective failures, with a threshold that changes at ten models, its error curve mixes group size with group composition. The view over all 48 pools applies no event threshold and leaves out no pool. Apart from the 523 groups of Section 2 that are degenerate or have no collective failure to measure against, it carries none of this selection, which is why it carries the headline reconstruction numbers.

Groups drawn from the strong-quarter candidate set of each pool (Table 6) show the same regularity as random groups: maximum entropy misses by 3.8% to 15.6% from three to eight models against 2.7% to 15.9%, with equal absolute errors. The strong groups have lower collective failure rates (median 14.8% against 23.4% at three models), which raises relative error at small sizes.

Table 3: **Relative reconstruction error for each pool in Figure 2.** Mean over groups, in percent of β.

		maximum entropy (% of β)					full-pair copula (% of β)				
Release	Pool	k=3	k=4	k=5	k=6	k=8	k=3	k=4	k=5	k=6	k=8
Co-failure release	ARC (A2)	3.9	10.0	7.1	12.6	7.7	64.6	73.5	85.7	82.3	97.8
	CodeGen (CG)	1.8	3.9	5.3	8.9	9.3	2.8	7.8	10.8	16.0	22.9
	GPQA (E2)	1.8	3.7	7.0	13.3	19.7	2.1	4.2	7.2	14.4	20.1
	GPQA-open	1.0	1.5	2.0	3.1	5.1	1.2	2.7	4.3	4.1	7.1
	GSM8K (A2)	8.1	11.1	15.3	13.5	3.1	47.7	76.3	83.6	92.0	99.1
	MATH-500 (E2)	2.0	3.3	4.3	5.5	5.9	2.6	4.8	7.8	7.7	12.1
	MATH-hard (MH)	3.7	5.1	7.3	9.0	11.8	5.0	9.2	15.3	21.3	28.3
	MMLU (A2)	7.4	14.4	20.3	27.8	37.6	11.3	23.3	34.2	42.1	58.1
	MMLU-Pro (E2)	3.9	8.1	9.3	12.7	18.4	4.4	10.2	12.0	16.0	24.9
LiveBench	Coding completion	1.7	3.7	5.3	6.4	12.3	2.6	4.8	7.4	10.8	17.2
	LCB generation	0.5	1.6	1.8	2.1	3.4	4.3	8.2	12.9	10.3	18.6
	Typos	2.7	5.9	6.9	10.1	12.8	2.7	5.9	7.3	10.6	13.8
HELM	AIR-Bench	1.2	1.6	2.7	3.2	5.5	0.9	1.9	3.0	3.2	3.8
	BBQ	3.8	7.4	16.7	25.2	64.2	8.1	12.7	17.8	120.3	109.1
	BoolQ	5.0	10.4	14.5	31.2	43.4	3.9	6.8	8.5	23.1	27.1
	CivilComments	3.7	9.9	22.2	25.8	58.2	4.8	11.4	30.1	31.3	62.6
	Commonsense	8.4	20.4	27.6	35.4	41.4	7.2	13.8	18.5	21.1	27.1
	Data imputation	2.1	3.6	4.6	3.7	4.5	2.8	7.5	9.9	17.1	27.4
	Dyck	2.3	7.5	14.1	21.1	63.9	2.3	7.1	13.1	20.1	59.2
	Entity matching	18.5	24.9	28.3	33.3	47.2	33.1	75.5	58.2	62.0	63.6
	GSM8K	2.6	5.9	8.5	10.7	10.3	2.9	7.1	10.5	16.1	21.2
	IMDB	7.1	15.0	24.1	21.4	37.5	4.7	10.4	10.5	14.2	18.7
	LSAT	0.6	1.3	2.1	2.5	6.8	0.9	1.3	2.8	3.0	9.8
	LegalBench	2.7	4.5	5.4	7.8	12.4	2.9	5.0	6.4	9.9	14.5
	LegalSupport	3.5	10.8	12.8	25.3	33.6	3.5	10.7	13.2	27.8	33.7
	MATH	2.4	5.5	7.6	10.5	14.4	2.9	6.0	9.6	13.2	22.6
	MMLU	4.4	9.4	14.1	18.3	23.5	2.1	4.0	5.8	7.9	12.0
	MedQA	2.3	3.0	6.0	8.2	10.6	1.7	2.4	4.9	5.2	7.9
	RAFT	3.2	6.1	7.3	10.2	10.7	3.3	6.5	7.9	10.5	10.7
	Synthetic reasoning	0.7	2.1	3.7	6.0	11.4	0.5	1.3	1.9	3.0	5.7
	ThaiExam	1.9	4.4	6.6	8.6	13.1	1.9	4.4	7.1	9.8	17.1
	TruthfulQA	0.7	1.5	2.3	3.5	6.2	0.7	1.7	2.7	4.3	8.2
	WikiFact	0.7	1.6	3.1	5.0	8.2	0.4	0.8	1.3	2.1	2.9
	bAbI	1.0	1.1	1.7	2.5	4.1	1.1	1.3	3.0	3.8	5.1
Open LLM Leaderboard	GPQA	1.4	3.0	5.5	6.7	9.5	1.6	3.5	7.7	9.8	16.8
	MATH-hard	0.4	1.2	2.3	3.9	7.1	0.3	0.5	1.0	1.0	1.6
	MMLU-Pro	1.0	1.7	1.9	1.9	1.5	1.1	2.8	3.9	5.3	8.0
	MuSR	0.9	1.4	2.7	4.4	8.3	0.8	1.0	1.8	2.2	2.5
EmbedLLM	ASDiv	0.0	0.0	0.1	0.3	0.5	0.0	0.2	0.3	0.4	0.8
	GPQA	0.6	1.6	3.2	5.1	9.3	0.4	0.9	1.7	2.4	4.6
	GSM8K	0.6	1.3	2.0	2.6	4.7	0.6	1.2	2.5	3.3	4.4
	LogiQA	0.8	1.4	1.9	2.6	4.9	0.8	1.3	1.8	2.2	3.6
	MMLU	2.4	3.9	6.9	7.9	10.9	2.1	3.4	6.5	7.6	12.4
	MathQA	0.3	0.8	1.8	3.1	7.0	0.4	1.0	1.4	1.9	2.9
	MedMCQA	1.0	1.8	3.0	4.7	6.7	1.1	2.1	4.0	6.7	10.6
	PIQA	1.0	1.4	1.7	2.4	4.4	2.1	4.0	6.7	8.9	11.6
	SocialIQA	0.6	1.5	1.7	2.7	3.3	15.3	25.7	39.4	46.3	59.4
	TruthfulQA	0.6	1.4	2.4	3.8	7.4	0.4	0.7	0.8	2.2	1.3

D ESTIMATORS

Every pairwise completion reads the same input: the k individual failure rates and the $\binom{k}{2}$ pairwise co-failure rates of the group. None sees item-level data. Direct counting reads the response matrix, and the interval midpoint is the centre of $[L, U]$.

Pairwise maximum entropy chooses the highest-entropy probability distribution that matches all supplied rates. The full-pair Gaussian copula gives every model a hidden Gaussian score and fits one correlation per pair (Emrich & Piedmonte, 1991; Sklar, 1959). Its equicorrelated version replaces these correlations with one shared value. The one- and two-factor probit models explain failures through one or two continuous hidden factors. The two-class latent model assumes that items come from two hidden classes and that models fail independently within each class. The Bahadur method truncates an expansion of the binary distribution after its pairwise terms, so it need not be a valid probability distribution (Bahadur, 1961). Independence simply multiplies the individual failure rates.

Table 4: **Release-level values behind Figure 2, in relative error and percentage points.** "Failures" in a row label is the median number of collective failures per group at eight models; the all-pool row pools the releases.

		relative error (% of β)					absolute error (percentage points)				
Release		k=3	k=4	k=5	k=6	k=8	k=3	k=4	k=5	k=6	k=8
Co-failure release	half-width of the compatible interval	12.9	24.6	34.6	43.5	50.6	1.31	2.23	3.11	3.74	3.93
9 pools, 67 entries, 8 failures	maximum entropy	3.7	6.8	8.6	11.8	13.2	0.33	0.49	0.59	0.79	0.84
	full-pair Gaussian copula	15.7	23.6	29.0	32.9	41.1	0.58	0.90	1.15	1.32	1.61
	independence	71.0	83.2	89.7	93.5	97.6	8.20	8.73	9.01	8.87	8.04
LiveBench	half-width of the compatible interval	6.3	14.4	25.6	37.7	49.3	1.55	3.01	4.48	5.49	6.04
3 pools, 150 entries, 10 failures	maximum entropy	1.6	3.7	4.7	6.2	9.5	0.43	0.84	0.94	0.98	1.25
	full-pair Gaussian copula	3.2	6.3	9.2	10.6	16.5	1.01	1.73	2.10	2.04	2.69
	independence	48.1	67.2	78.6	86.0	95.2	15.21	17.64	18.14	17.81	15.33
HELM	half-width of the compatible interval	23.2	54.3	85.1	113.9	189.7	2.59	4.62	5.86	6.46	6.89
22 pools, 183 entries, 43 failures	maximum entropy	3.6	7.2	10.7	14.5	24.1	0.33	0.50	0.64	0.74	0.92
	full-pair Gaussian copula	4.2	9.1	11.2	19.5	25.9	0.38	0.50	0.62	0.74	0.85
	independence	59.2	72.2	77.2	86.3	90.6	7.52	7.93	8.01	7.75	7.36
Open LLM Leaderboard	half-width of the compatible interval	10.1	23.4	40.1	56.9	78.3	3.83	7.66	11.44	14.56	17.33
4 pools, 346 entries, 382 failures	maximum entropy	0.9	1.8	3.1	4.2	6.6	0.35	0.61	0.96	1.21	1.70
	full-pair Gaussian copula	0.9	2.0	3.6	4.6	7.2	0.36	0.63	1.00	1.08	1.37
	independence	36.6	51.2	62.3	71.7	82.8	14.07	17.69	19.51	20.45	20.67
EmbedLLM	half-width of the compatible interval	8.6	21.1	35.9	48.0	66.3	2.49	5.09	7.51	9.02	10.44
10 pools, 112 entries, 488 failures	maximum entropy	0.8	1.5	2.5	3.5	5.9	0.25	0.45	0.70	0.98	1.53
	full-pair Gaussian copula	2.3	4.0	6.5	8.2	11.2	1.00	1.46	2.16	2.29	2.68
	independence	34.6	47.8	58.0	65.1	74.5	10.65	13.17	15.26	15.59	15.96
All 48 pools	half-width of the compatible interval	**16.1**	**36.8**	**57.9**	**77.5**	**119.9**	**2.37**	**4.42**	**6.07**	**7.10**	**7.89**
48 pools, 858 entries	maximum entropy	**2.7**	**5.3**	**7.6**	**10.3**	**15.9**	**0.32**	**0.52**	**0.69**	**0.85**	**1.12**
	full-pair Gaussian copula	**5.6**	**10.0**	**12.8**	**17.9**	**23.6**	**0.59**	**0.86**	**1.16**	**1.28**	**1.53**
	independence	**53.7**	**67.1**	**74.4**	**82.0**	**88.2**	**9.33**	**10.59**	**11.30**	**11.28**	**10.89**

Table 5: **The balanced design excludes nine small or low-failure pools.** Entries count qualifying groups among 400 candidates; values below the required 30 are bold.

					groups passing, of 400				
Release	Pool	Items	Models	Err	k=3	k=4	k=5	k=6	k=8
Co-failure release	ARC (A2)	120	15	6%	**0**	**0**	**0**	**0**	**0**
Co-failure release	CodeGen (CG)	63	18	56%	132	74	**20**	**6**	**0**
Co-failure release	GSM8K (A2)	120	15	7%	**0**	**0**	**0**	**0**	**0**
Co-failure release	MMLU (A2)	120	15	24%	**5**	**0**	**0**	**0**	**0**
Co-failure release	MMLU-Pro (E2)	200	52	35%	218	105	42	**6**	**0**
LiveBench	Coding completion	50	149	64%	156	71	48	**28**	**3**
LiveBench	Typos	100	58	47%	309	189	94	44	**3**
HELM	Commonsense	498	91	17%	84	**17**	**4**	**0**	**0**
HELM	Entity matching	1,396	67	26%	214	126	65	47	**17**

We fit maximum entropy by optimising a convex objective equivalent to entropy maximisation. The Gaussian methods compute multivariate normal probabilities (Genz, 1992); the factor models use numerical quadrature; and the latent-class and Bahadur methods use least squares. If the copula

Table 6: **Reconstruction error on strong-quarter groups against random groups.** Strong-quarter groups are drawn from the $\max(\lceil M/4 \rceil, 2k)$ most accurate candidates of each pool (60 per pool and size, or all when fewer exist); random groups are the main-analysis groups of Table 1. Relative error in percent of β and absolute error in percentage points, averaged within each pool and then across pools. $\tilde{\beta}$: median collective failure rate of the groups, in percent.

	strong-quarter groups					random groups				
k	$\tilde{\beta}$	max. ent. (%)	copula (%)	max. ent. (pp)	copula (pp)	$\tilde{\beta}$	max. ent. (%)	copula (%)	max. ent. (pp)	copula (pp)
3	14.8	3.8	7.4	0.36	0.44	23.4	2.7	5.6	0.32	0.59
4	12.0	6.5	10.0	0.53	0.63	16.5	5.3	10.0	0.52	0.86
5	10.3	9.5	12.3	0.69	0.74	13.5	7.6	12.8	0.69	1.16
6	9.3	13.0	15.5	0.86	0.93	11.2	10.3	17.9	0.85	1.28
8	8.1	15.6	19.6	1.09	1.13	9.0	15.9	23.6	1.12	1.53

correlation matrix is not positive definite, we replace its negative eigenvalues with small positive values and rescale its diagonal. This repair makes the calculation possible, but it also means the fitted copula no longer reproduces the original pairwise rates.

Error metrics. Let $\hat{\beta}$ denote an estimate of the true collective failure rate β. Relative error is $|\hat{\beta} - \beta|/\beta$; absolute error is the same gap in percentage points. Relative error prevents pools with common collective failure from dominating the average, while absolute error shows the operational size of a miss. We report both. They rank maximum entropy and the full-pair copula the same within all but one pool at eight models, and that exception differs by less than one hundredth of a percentage point. When results are averaged across pools with different values of β, however, the metrics need not agree. At eight models, maximum entropy has lower mean relative error than the copula, while their mean absolute errors are not clearly separated. We therefore do not use either averaged metric alone to rank the two completions.

The dependence on scale is substantial. For β in $(0, 1\%]$, $(1, 5\%]$, $(5, 20\%]$, and above 20%, maximum entropy's mean relative error at eight models is 61.6%, 19.4%, 8.6%, and 6.6%, respectively. Mean absolute error moves in the opposite direction: 0.23, 0.56, 0.85, and 2.10 percentage points. This is why we report both metrics and why the held-out analysis is organised by the number of observed collective failures. Table 7 gives signed and tail errors. Maximum entropy's median signed error is zero through five models and -0.14 points (-1.8% of β) at eight, where it lies below β on 59% of groups. The copula's is -0.23 points (-2.5%), below on 60%. At eight models the 95th percentile of relative error is 49.8% for maximum entropy and 99.4% for the copula.

Table 7: **Signed and tail reconstruction errors.** Main-analysis groups of Table 1. MAE: mean absolute error in percentage points, averaged within each pool and then across pools. Below: share of groups whose estimate lies below β. Median: median signed error over all groups at that size, in points and in percent of β. P90 and P95: percentiles of relative (percent of β) and absolute (points) error over all groups.

k	Completion	Groups	MAE (pp)	Below (%)	Median (pp)	Median (%)	P90 (%)	P95 (%)	P90 (pp)	P95 (pp)
3	Maximum entropy	2837	0.32	48.2	+0.00	+0.0	6.2	9.8	0.72	0.95
3	Full-pair copula	2837	0.59	53.7	-0.02	-0.1	9.5	22.7	0.89	1.36
4	Maximum entropy	2801	0.52	51.9	-0.00	-0.0	12.2	20.1	1.18	1.45
4	Full-pair copula	2801	0.86	57.8	-0.11	-0.5	21.7	47.6	1.48	2.18
5	Maximum entropy	2789	0.69	51.2	-0.00	-0.0	17.9	27.9	1.65	2.06
5	Full-pair copula	2789	1.16	56.4	-0.13	-0.8	32.8	69.3	1.92	2.79
6	Maximum entropy	2740	0.85	54.8	-0.06	-0.5	24.3	37.7	2.16	2.83
6	Full-pair copula	2740	1.28	59.0	-0.18	-1.5	45.2	76.6	2.21	3.22
8	Maximum entropy	2710	1.12	59.0	-0.14	-1.8	32.2	49.8	2.86	4.10
8	Full-pair copula	2710	1.53	59.8	-0.23	-2.5	62.0	99.4	2.96	4.13

The main relative-error analysis omits 523 groups for which relative error is undefined or a marginal is degenerate. The 81 with $\beta = 0$ include every group with a member that always succeeds. The other 442 contain a member that always fails. Fitting maximum entropy on the corresponding boundary face gives 0.40 percentage points mean absolute error over all 523 groups and 0.60 points at eight models. Their omission therefore affects the relative-error estimand but does not hide a large absolute-error failure.

EmbedLLM's public matrix includes repeated rows and, for GPQA, several prompt variants of overlapping questions. Treating its ten benchmarks as ten pools follows the release, but the rows are not all independent benchmark questions. Its low released scores on ASDiv, GPQA, and SocialIQA are therefore not interpreted as clean estimates of benchmark ability: ASDiv is open-ended rather than a four-choice task, and the GPQA row count reflects prompt variants rather than 5,960 distinct questions. Excluding EmbedLLM leaves 38 pools: at eight models, the half-width of the compatible interval is 134% of β, while maximum entropy, the full-pair copula, and independence miss by 18.5%, 26.8%, and 91.8%. The headline ordering is therefore not created by that release.

D.1 Maximum entropy and the Gaussian copula

Maximum entropy reproduces the input. Across all 13,877 reconstruction groups, the largest mismatch between a supplied rate and the maximum-entropy fit has a median of 1.3×10^{-9}. Only 37 near-boundary cases exceed 2×10^{-6}, and the worst mismatch is 0.11 percentage points. Every

fit is finite. By contrast, the full-pair copula needs its correlation matrix repaired for 30% of eight-model groups. On those groups, its median largest mismatch with a supplied pairwise rate is 1.9 percentage points.

The methods separate when the copula needs repair. When the copula matrix is valid, maximum entropy and the copula have similar error. At eight models the copula is slightly ahead. When repair is needed, maximum entropy is more accurate on 82% of groups and its mean error is less than half as large (Table 8). This explains most of the overall gap between the methods.

Table 8: **Maximum entropy and the copula are similar when the copula is valid; repair separates them.** Values are relative error averaged over groups.

Copula matrix	Group size	Groups	Max. entropy / copula	Max. entropy lower
Valid	8	1,890	15.7 / 14.5	48%
Repaired	8	820	15.4 / 43.6	82%
Valid	3–8	11,439	7.9 / 8.8	50%
Repaired	3–8	2,438	9.7 / 37.6	83%

The original repair clips negative eigenvalues and rescales the diagonal. We also tested Higham's alternating projection to the nearest correlation matrix (Higham, 2002). On the 820 groups requiring repair, averaged within each pool and then across the 32 pools that contain them, relative error is 15.5% for maximum entropy, 30.9% after eigenvalue clipping, and 30.0% after nearest-correlation projection (Table 8 averages over groups). The projection still misses at least one supplied pairwise rate by a median of 1.63 percentage points per group. The copula result is therefore not an artefact of choosing the cruder repair.

The conclusion is not uniform across pools. At eight models, maximum entropy has lower mean relative error on every release (the co-failure release, LiveBench, HELM, the Open LLM Leaderboard, and EmbedLLM), but not on every pool (Table 3). The difference is significant on the first two and uncertain on the remaining three. It is not significantly worse than the copula in either relative error or percentage points on any release. Four pools—ARC and GSM8K from the co-failure release, HELM BBQ, and EmbedLLM SocialIQA—account for 78% of the signed average gap. The first two contain only 120 items and roughly one collective failure per sampled group. Table 4 reports the complete release-level results.

Exact fitting has a clear size limit. The implementation enumerates all 2^k failure patterns, so time and memory grow exponentially. In a single-process timing run, fitting takes 1.6 seconds at 16 models, 14 seconds at 18, and about three minutes with 4.2 GB of memory at 20. We use exact fitting only. Groups much larger than 20 would need an approximate method, which we have not evaluated.

E STATISTICAL PROCEDURE

Groups drawn from one pool share both models and items, so they are not independent observations. Every average is taken within a pool first and then across pools. Descriptive reconstruction intervals bootstrap the 48 pools with 4,000 resamples. Comparisons are paired on the same groups before resampling.

The choice matters. Reconstruction error at $k{=}8$ is 15.9% with a pool-level interval of $[11.6, 20.7]$. Bootstrapping the same numbers over the 2,710 individual groups instead gives $[14.3, 17.0]$, an interval 3.3 times narrower and centred on the group mean rather than the pool mean, because it treats the groups drawn from one pool, between 21 and 60 of them at this size, as that many independent facts. Several comparisons reported in this paper as ties would appear significant under that treatment.

F DEPENDENCE BY INTERACTION ORDER

Estimating a distribution over 2^k outcomes needs many items, so the cohorts are fixed in advance by a sample-size rule of at least $40 \cdot 2^k$ items, which depends on group size and not on any result. Two cohorts hold the pools and the item count fixed as k grows. The first is the 13 pools that reach six models, each group scored on 2,560 items: four EmbedLLM pools, eight HELM pools, and MMLU-Pro from the Open LLM Leaderboard. The second is the 3 pools large enough for eight models, at 10,240 items. For a group of k models, let P_j be the maximum-entropy distribution matching all observed failure moments through order j, and let S_j be its entropy. The empirical distribution is P_k. Then $D_{\mathrm{KL}}(P_k \,\|\, P_j) = S_j - S_k$, and the share carried by order two is $(S_1 - S_2)/(S_1 - S_k)$. Pinsker's inequality gives $\mathrm{TV}(P_k, P_2) \le \sqrt{D_{\mathrm{KL}}(P_k \,\|\, P_2)/2}$, and $|P_k(A) - P_2(A)| \le \mathrm{TV}$ for A the all-fail event, so $|\beta - \hat{\beta}_2| \le \sqrt{(S_2 - S_k)/2}$ with the entropies in nats. On the six-model cohort this bound averages 10.4 percentage points against an observed miss of 1.1, and on the eight-model cohort 13.8 against 0.7. It holds on every group.

At six models the share carried by order two runs from 88% to 98% across the 13 pools, and 12 of the 13 sit above 94%. Applying the sample-size rule directly instead of a fixed item count gives the same share at six models. Using all available items instead of a uniform count moves the six-model share by less than one point. The Miller–Madow (Miller, 1955) and Grassberger (Grassberger, 2003) corrections both raise the shares relative to the plug-in estimate, so the plug-in values reported in Section 4 are the conservative ones. This finite-k decomposition is descriptive. A strong pairwise share for the group sizes measured here does not imply that the share remains high as the number of models grows (Roudi et al., 2009). Table 9 lists every pool of both cohorts.

Table 9: **Share of measured dependence carried by the pairwise terms, pool by pool.** Mean over the groups of each pool of $(S_1 - S_2)/(S_1 - S_k)$ in percent, plug-in entropy and the Grassberger-corrected estimate. Six-model cohort: 13 pools at 2,560 items. Eight-model cohort: 3 pools at 10,240 items. The correction is applied to each entropy separately, so a corrected share can exceed 100% slightly (MathQA).

Pool (k=6)	plug-in	corrected	Pool (k=8)	plug-in	corrected
MathQA (EmbedLLM)	98.4	100.1	MMLU-Pro (Open LLM Leaderboard)	98.3	99.1
MMLU-Pro (Open LLM Leaderboard)	98.4	99.6	MMLU (HELM)	95.8	96.9
AIR-Bench (HELM)	98.1	99.6	MMLU (EmbedLLM)	95.6	96.9
MedMCQA (EmbedLLM)	96.5	99.9			
MMLU (HELM)	96.4	98.1			
MMLU (EmbedLLM)	96.3	98.5			
WikiFact (HELM)	96.2	97.6			
BoolQ (HELM)	95.9	98.6			
IMDB (HELM)	95.8	98.3			
bAbI (HELM)	94.3	97.2			
CivilComments (HELM)	94.3	96.7			
Synthetic reasoning (HELM)	94.3	96.7			
GPQA (EmbedLLM)	88.4	91.8			

G CHOOSING A GROUP

Steps. For each pool, the items are permuted with a fixed seed and split into two halves. Each half in turn is the fit half. On the fit half we compute each model's failure rate p_i and each pair's co-failure rate p_{ij}. The candidates are the models with $0 < p_i < 1$ on the fit half. A group is built one model at a time: starting from the empty group, the candidate whose addition gives the lowest estimated collective failure rate is added, until the group has k members. Ties keep the lower column index. The chosen group is then scored by the fraction of items in the other half on which all its members fail. Split directions on which the reference group has no collective failure on the scored half are excluded, which removes 2, 4, and 6 of the 96 at $k = 3$, 5, and 8. The reduction is $1 - \beta_{\text{chosen}}/\beta_{\text{reference}}$, averaged over the two directions within a pool and then over pools, with a 4,000-resample bootstrap over pools.

Selectors. *Accuracy ranking* takes the k candidates with the lowest failure rates. *Maximum entropy* estimates a group's collective failure rate by the maximum-entropy distribution matching the group's individual and pairwise rates. The *full-pair copula* estimates it by a Gaussian copula with the tetra-

choric correlation of each pair, repaired when the matrix is not positive definite. The *equicorrelated copula* uses one shared correlation, the mean of the pairwise values. *Maximum disagreement* adds the candidate that gives the highest mean pairwise disagreement rate $p_i + p_j - 2p_{ij}$ (Kuncheva & Whitaker, 2003), starting from the most accurate candidate. Every selector sees the same candidates and the same fit-half rates.

Settings. *Whole pool*: every candidate. *Strong quarter*: the candidates are cut to the $\max(\lceil M/4 \rceil, 2k)$ most accurate on the fit half, where M is the number of candidates. The *strong half* uses $\max(\lceil M/2 \rceil, 2k)$. The rule was fixed before the comparison was run. Split directions with too few candidates for a cut to remove anything are left out of that setting, which removes one pool at $k = 5$ and four at $k = 8$, so the strong-quarter setting has 48, 47, and 43 pools at $k = 3$, 5, and 8, against 48, 48, and 47 for the whole pool, where one pool has no collective failure of the reference on either scored half at $k = 8$. Accuracy ranking is unchanged by either cut, because its members already lie in the retained set. *One-per-family reference*: the reference walks down the fit-half accuracy ranking and takes a model only if its family is not yet in the group. When a full pass fills fewer than k slots, a family may appear twice, which happens on 5% of the $k = 8$ split directions. Table 10 reports every selector in the four settings, and Figure 7 shows the same comparison pool by pool.

Table 11 gives the same comparison in percentage points with no split direction excluded: on the whole pool the maximum-entropy group's held-out collective failure is 2.8, 3.4, and 3.3 points below accuracy ranking at three, five, and eight models, and under the strong-quarter rule 2.1, 2.1, and 2.0 points, with every interval below zero.

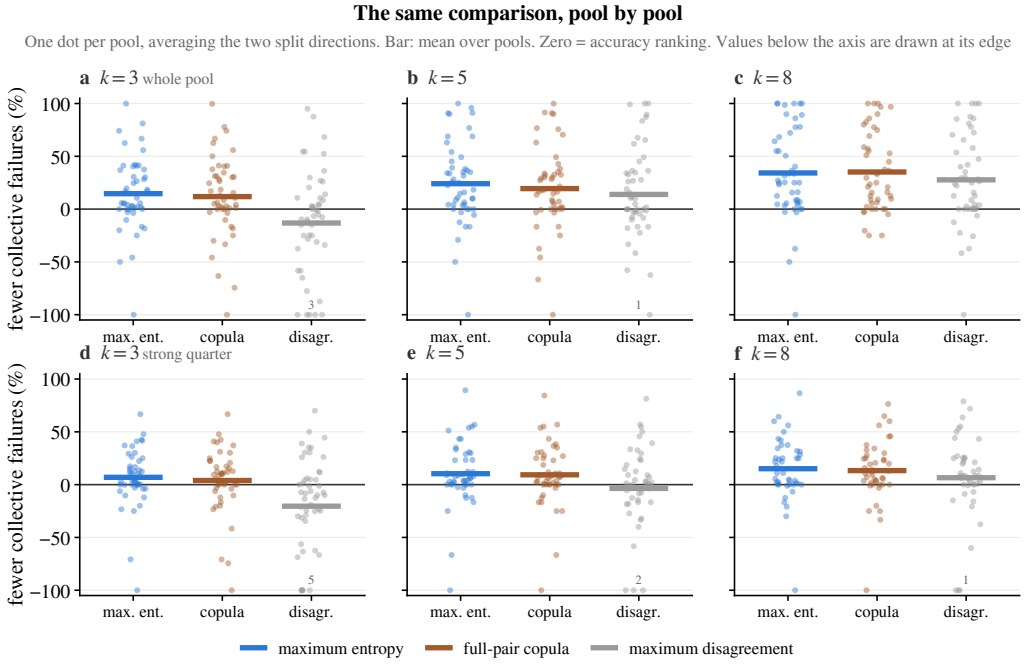


Figure 7: **The 48-pool comparison, pool by pool.** Each dot is one pool, averaged over the two split directions. The bar is the mean over pools. Values below -100% are drawn at the axis edge with their count. Top row: whole pool, 48 pools at $k = 3$ and 5 and 47 at $k = 8$. Bottom row: candidates under the strong-quarter rule (top quarter, $2k$ floor), 48, 47, and 43 pools. The reference is accuracy ranking.

Families. For open-weight models, the family is the architecture class of the weights as recorded by the Open LLM Leaderboard metadata or, for EmbedLLM models absent from it, by the model's configuration file on the Hub. Versions of one line share a label (Qwen2, Llama, Gemma, Mistral and Mixtral, Phi). Models with the Llama architecture, including the Yi models, therefore share the Llama label. For API models, the family is the lab that pretrained the base weights, and a fine-tune that names its base inherits the base's label. Four models that could not be resolved received

Table 10: **Every selector in the four 48-pool settings.** Reduction: mean relative reduction over pools in held-out collective failure against the reference, in percent, with the 95% pool-level bootstrap interval. Ratio: median over split directions of the chosen group's rate over the reference's. Won / tied: split directions on which the selector's group fails together less often / equally often. Pools: pools on which its mean is lower / higher / equal. Accuracy: mean held-out accuracy of the chosen members. Families: distinct families in the chosen group. Split directions on which the reference has no collective failure on the scored half are excluded, and the quarter and half cuts leave out pools too small to cut, so the pool count, the reference accuracy, and the family counts differ slightly between settings. A dash marks a quantity not recorded for that selector.

Setting	Selector	k	Reduction [95% CI]	Ratio	Won / tied	Pools −/+/=	Accuracy	Families
Whole pool, reference accuracy ranking		3	reference	1.00	–	–	71.0	2.1
	Maximum entropy	3	14.6 [5.1, 24.3]	0.88	66 / 12	35 / 9 / 4	61.8	2.5
	Full-pair copula	3	11.8 [1.1, 22.5]	0.90	64 / 12	33 / 11 / 4	62.2	2.5
	Equicorrelated copula	3	6.2 [-3.2, 15.3]	0.95	58 / 12	29 / 15 / 4	–	2.6
	Maximum disagreement	3	-13.2 [-28.1, 1.7]	1.03	42 / 3	20 / 26 / 2	–	2.6
		5	reference	1.00	–	–	69.4	3.0
	Maximum entropy	5	24.1 [13.1, 35.1]	0.77	70 / 12	36 / 6 / 6	57.4	3.7
	Full-pair copula	5	19.4 [8.0, 30.6]	0.81	68 / 10	34 / 10 / 4	57.7	3.8
	Equicorrelated copula	5	14.2 [3.1, 25.0]	0.83	59 / 11	29 / 15 / 4	–	3.7
	Maximum disagreement	5	13.9 [-1.0, 27.2]	0.86	54 / 9	27 / 16 / 5	–	3.8
		8	reference	1.00	–	–	67.4	4.1
	Maximum entropy	8	34.2 [21.4, 46.1]	0.69	71 / 12	39 / 5 / 3	54.8	5.0
	Full-pair copula	8	35.1 [24.9, 45.4]	0.71	70 / 12	36 / 6 / 5	53.8	5.0
	Equicorrelated copula	8	24.2 [12.7, 35.5]	0.75	63 / 10	31 / 10 / 6	–	5.0
	Maximum disagreement	8	27.7 [15.8, 39.9]	0.75	64 / 13	33 / 7 / 7	–	5.2
Strong quarter, reference accuracy ranking		3	reference	1.00	–	–	71.0	2.1
	Maximum entropy	3	6.9 [-0.9, 14.3]	0.95	57 / 21	31 / 11 / 6	67.8	2.5
	Full-pair copula	3	3.9 [-4.5, 12.0]	0.95	56 / 19	32 / 12 / 4	67.9	2.5
	Equicorrelated copula	3	4.5 [-2.8, 11.2]	0.98	54 / 20	31 / 12 / 5	68.0	–
	Maximum disagreement	3	-20.4 [-36.5, -6.1]	1.04	34 / 7	18 / 27 / 3	62.9	–
		5	reference	1.00	–	–	69.5	3.0
	Maximum entropy	5	10.4 [1.5, 18.6]	0.92	61 / 17	33 / 8 / 6	65.2	3.6
	Full-pair copula	5	9.3 [0.9, 17.4]	0.93	60 / 18	32 / 9 / 6	65.4	3.6
	Equicorrelated copula	5	8.3 [0.3, 15.5]	0.92	62 / 14	34 / 9 / 4	65.7	–
	Maximum disagreement	5	-3.6 [-17.1, 8.3]	0.94	53 / 4	23 / 20 / 4	61.8	–
		8	reference	1.00	–	–	66.2	4.0
	Maximum entropy	8	15.1 [5.6, 23.6]	0.86	58 / 15	32 / 6 / 5	62.3	4.8
	Full-pair copula	8	13.3 [4.1, 21.7]	0.89	59 / 15	32 / 6 / 5	62.9	4.7
	Equicorrelated copula	8	14.4 [8.1, 20.4]	0.89	60 / 14	32 / 5 / 6	63.3	–
	Maximum disagreement	8	6.6 [-5.6, 17.6]	0.92	54 / 11	31 / 10 / 2	60.8	–
Strong half, reference accuracy ranking		3	reference	1.00	–	–	71.0	2.1
	Maximum entropy	3	10.9 [1.9, 19.2]	0.92	60 / 18	33 / 9 / 6	66.5	2.6
	Full-pair copula	3	6.3 [-3.7, 16.2]	0.92	59 / 17	31 / 12 / 5	66.6	2.5
	Equicorrelated copula	3	7.1 [-1.6, 15.1]	0.95	57 / 17	30 / 13 / 5	66.4	–
	Maximum disagreement	3	-31.1 [-54.5, -9.8]	1.07	33 / 3	16 / 31 / 1	58.8	–
		5	reference	1.00	–	–	69.5	3.0
	Maximum entropy	5	18.5 [10.8, 26.6]	0.85	65 / 15	34 / 6 / 7	62.8	3.7
	Full-pair copula	5	13.9 [3.9, 23.4]	0.86	67 / 11	35 / 8 / 4	63.4	3.7
	Equicorrelated copula	5	13.1 [3.6, 21.8]	0.86	64 / 12	33 / 9 / 5	63.5	–
	Maximum disagreement	5	-19.2 [-55.3, 6.8]	0.98	48 / 7	25 / 19 / 3	57.7	–
		8	reference	1.00	–	–	66.2	4.0
	Maximum entropy	8	25.8 [16.7, 35.1]	0.82	68 / 10	38 / 2 / 3	59.5	5.1
	Full-pair copula	8	21.0 [11.7, 30.4]	0.83	65 / 13	35 / 4 / 4	60.3	4.9
	Equicorrelated copula	8	20.2 [11.5, 29.1]	0.87	61 / 16	34 / 5 / 4	60.7	–
	Maximum disagreement	8	0.6 [-25.5, 21.8]	0.83	61 / 4	31 / 11 / 1	56.5	–
Whole pool, reference one-per-family ranking		3	reference	1.00	–	–	69.7	3.0
	Maximum entropy	3	8.1 [-5.1, 19.8]	0.94	61 / 16	31 / 8 / 9	61.8	2.5
	Full-pair copula	3	4.2 [-11.6, 16.9]	0.95	61 / 13	31 / 10 / 7	62.2	2.5
	Equicorrelated copula	3	-2.4 [-19.2, 10.5]	0.97	51 / 13	27 / 14 / 7	–	2.6
	Maximum disagreement	3	-23.2 [-44.8, -5.2]	1.07	35 / 2	15 / 31 / 2	–	2.6
	One-per-family vs. accuracy ranking	3	3.2 [-3.0, 8.9]	1.00	38 / 36	20 / 12 / 16	–	–
		5	reference	1.00	–	–	66.1	5.0
	Maximum entropy	5	19.1 [4.1, 32.3]	0.80	73 / 9	37 / 5 / 6	56.8	3.8
	Full-pair copula	5	14.5 [0.3, 27.3]	0.83	67 / 12	35 / 8 / 5	57.1	3.8
	Equicorrelated copula	5	15.3 [4.5, 26.1]	0.89	59 / 9	32 / 12 / 4	–	3.7
	Maximum disagreement	5	11.0 [-4.9, 24.6]	0.88	59 / 8	28 / 13 / 7	–	3.8
	One-per-family vs. accuracy ranking	5	4.3 [-0.5, 10.2]	1.00	38 / 25	23 / 17 / 8	–	–
		8	reference	1.00	–	–	62.1	7.9
	Maximum entropy	8	27.2 [11.3, 41.7]	0.78	74 / 10	39 / 6 / 3	54.4	5.0
	Full-pair copula	8	28.1 [13.8, 40.7]	0.76	71 / 13	38 / 5 / 5	53.2	5.0
	Equicorrelated copula	8	23.3 [12.3, 34.5]	0.82	67 / 12	34 / 8 / 6	–	5.1
	Maximum disagreement	8	24.5 [12.4, 36.6]	0.81	66 / 12	33 / 6 / 9	–	5.3
	One-per-family vs. accuracy ranking	8	4.4 [-1.1, 10.2]	1.00	40 / 22	21 / 16 / 10	–	–

Table 11: **Group choice in percentage points.** Held-out collective failure of the chosen group minus that of accuracy ranking, in points, averaged over the two split directions of each pool and then over pools, with a 95% pool-level bootstrap interval; no split direction is excluded, not even those on which the reference group shows no collective failure. Negative values mean fewer collective failures. Directions: split directions on which the chosen group fails together less often / equally often / more often than accuracy ranking. Reference: mean held-out collective failure of accuracy ranking, in percent.

Setting	Selector	k	Pools	Reference (%)	Difference (pp) [95% CI]	Directions −/=/+
Whole pool	Maximum entropy	3	48	16.91	-2.85 [-4.56, -1.41]	66 / 13 / 17
	Full-pair copula	3	48	16.91	-2.55 [-4.24, -1.13]	64 / 14 / 18
	Equicorrelated copula	3	48	16.91	-1.76 [-3.05, -0.58]	58 / 14 / 24
	Maximum disagreement	3	48	16.91	-0.43 [-2.01, +1.04]	42 / 4 / 50
	Maximum entropy	5	48	13.70	-3.44 [-5.24, -1.82]	70 / 13 / 13
	Full-pair copula	5	48	13.70	-2.85 [-4.25, -1.61]	68 / 12 / 16
	Equicorrelated copula	5	48	13.70	-2.39 [-3.55, -1.33]	59 / 14 / 23
	Maximum disagreement	5	48	13.70	-2.36 [-3.96, -0.90]	54 / 11 / 31
	Maximum entropy	8	48	11.46	-3.26 [-4.89, -1.84]	71 / 13 / 12
	Full-pair copula	8	48	11.46	-3.10 [-4.52, -1.81]	70 / 14 / 12
	Equicorrelated copula	8	48	11.46	-2.45 [-3.44, -1.50]	63 / 13 / 20
	Maximum disagreement	8	48	11.46	-2.89 [-4.32, -1.68]	64 / 15 / 17
Strong quarter	Maximum entropy	3	48	16.91	-2.11 [-3.64, -0.84]	57 / 23 / 16
	Full-pair copula	3	48	16.91	-1.82 [-3.30, -0.61]	56 / 21 / 19
	Equicorrelated copula	3	48	16.91	-1.50 [-2.63, -0.45]	54 / 22 / 20
	Maximum disagreement	3	48	16.91	-0.04 [-1.85, +1.50]	34 / 9 / 53
	Maximum entropy	5	48	13.70	-2.10 [-3.44, -0.98]	61 / 19 / 15
	Full-pair copula	5	48	13.70	-1.99 [-3.20, -0.93]	60 / 20 / 15
	Equicorrelated copula	5	48	13.70	-1.78 [-2.66, -0.99]	62 / 17 / 16
	Maximum disagreement	5	48	13.70	-1.30 [-2.64, -0.17]	53 / 6 / 36
	Maximum entropy	8	44	11.81	-1.98 [-3.02, -1.03]	58 / 17 / 13
	Full-pair copula	8	44	11.81	-1.94 [-2.92, -1.06]	59 / 18 / 11
	Equicorrelated copula	8	44	11.81	-1.83 [-2.68, -1.09]	60 / 17 / 11
	Maximum disagreement	8	44	11.81	-1.63 [-2.72, -0.64]	54 / 14 / 20

singleton labels, which can only make the one-per-family reference easier to satisfy. The 48 pools contain between 5 and 20 families each. At $k = 8$, the accuracy-ranked group spans 4.1 families on average, the one-per-family reference 7.9, and the maximum-entropy group 5.0.

Worked examples. On Open LLM Leaderboard MMLU-Pro, the five models most accurate on the fit half are three calme-78B variants, CalmeRys-78B, and Reflection-Llama-70B. They represent two families and fail together on 1,042 of the 5,997 scored-half items (17.4%). The one-per-family reference gives calme-3.2-78B, Reflection-Llama-70B, Gemma2atlas-27B, Chocolatine-14B, and Mistral-Large-2411, which fail together on 721 items (12.0%). Maximum entropy selects calme-3.2-78B, QwQ-32B, Reflection-Llama-70B, Mistral-Large-2411, and Qwen2.5-Math-14B, which fail together on 672 items (11.2%). The strong-quarter cut selects the same group. On HELM MMLU at five models, accuracy ranking on the fit half takes two Claude 3.5 Sonnet versions, Llama 3.1 405B, GPT-4o, and Gemini 1.5 Pro, which fail together on 4.8% of the 7,021 scored-half items with a mean member accuracy of 86.2%. Unrestricted maximum entropy takes Claude 3.5 Sonnet, Command R, two OLMo-7B versions, and Gemini 1.5 Pro, which fail together on 3.4% with a mean member accuracy of 63.8%. Under the strong-quarter cut it takes Yi-Large, Palmyra-X-004, Claude 3.5 Sonnet, GPT-4o, and Gemini 1.5 Pro, which fail together on 4.4% with a mean member accuracy of 84.0%.

Two further collections. *SWE-bench Verified.* The 164 systems of Appendix H over the 500 tasks. Twenty random splits of the tasks, both directions, with the steps above unchanged. The interval resamples the twenty splits. The one-agent subset holds the 41 systems that run mini-SWE-agent with different models. The splits share the same 500 tasks and are not independent samples. *RewardBench 2, coverage.* The 165 reward models with scalar per-response scores in the public results, on the five standard skills (the Ties skill is excluded, as in the majority-vote analysis). Every prompt is expanded into its chosen-versus-rejected response pairs, and a judge fails a pair when it scores the rejected response at least as high as the chosen one. Each skill's prompts are split in half at random, so every pair of a prompt falls on the same side, in both directions, which gives ten split directions, of which the ones where accuracy ranking has no collective failure are excluded (two at each k). The interval resamples split directions. A prompt-cluster bootstrap that resamples the scored-half prompts of every split direction gives intervals of similar width (for maximum entropy,

[86, 94], [97, 100], and [97, 100] at three, five, and eight judges). *RewardBench 2, majority vote.* One of the same five skills is held out in turn. Individual and pairwise judge errors are computed on the other four, with equal weight per skill, per prompt within a skill, and per pair within a prompt. Maximum entropy scores 300 random committees by their estimated rate of a wrong majority, the probability under the maximum-entropy distribution that at least $\lceil k/2 \rceil$ judges err, and then improves the best by single-member swaps, once. The committee is then frozen and its majority decisions are evaluated on the held-out skill. Strict prompt accuracy requires the majority to rank the chosen response above each of its three distractors. Committee sizes are 3, 5, and 7. The interval against the top-judge committee is a prompt-cluster bootstrap. Table 12 gives the results.

Table 12: **The two further collections.** SWE-bench Verified: 20 random task splits, both directions, and the interval resamples splits. RewardBench 2 coverage: the prompts of each skill split in half, both directions, and the interval resamples split directions. "Events" is the median, over all ten split directions, of the number of held-out comparisons on which the whole panel fails. RewardBench 2 majority vote: leave one skill out, strict prompt accuracy of the committee's majority decision on the held-out skill, against the top-judge committee, the majority vote of the k judges with the highest accuracy on the source skills. The interval for maximum entropy is a prompt-cluster bootstrap. Gains are computed from unrounded accuracies.

Collection	Selector	k	Reduction [95% CI]	Won / tied	Events
SWE-bench Verified, 164 systems	Maximum entropy	3	11.4 [7.9, 15.0]	31 / 2	–
	Full-pair copula	3	0.2 [-6.0, 5.8]	23 / 2	–
	Equicorrelated copula	3	-7.3 [-11.8, -2.6]	12 / 4	–
	Maximum disagreement	3	-10.4 [-15.4, -5.6]	8 / 3	–
	Maximum entropy	5	15.4 [11.7, 18.9]	32 / 4	–
	Full-pair copula	5	-1.2 [-8.9, 5.3]	22 / 4	–
	Equicorrelated copula	5	-9.8 [-14.5, -5.2]	7 / 3	–
	Maximum disagreement	5	-8.1 [-12.0, -4.3]	11 / 4	–
	Maximum entropy	8	9.3 [6.7, 11.7]	31 / 3	–
	Full-pair copula	8	-0.9 [-5.1, 3.2]	22 / 1	–
	Equicorrelated copula	8	-16.4 [-20.7, -12.1]	3 / 1	–
	Maximum disagreement	8	-9.9 [-13.1, -6.8]	7 / 1	–
SWE-bench Verified, one agent, 41 systems	Maximum entropy	3	7.3 [5.3, 9.2]	32 / 2	–
	Full-pair copula	3	6.0 [3.6, 8.3]	30 / 2	–
	Equicorrelated copula	3	-3.5 [-6.0, -1.1]	10 / 6	–
	Maximum disagreement	3	-31.9 [-38.3, -26.1]	4 / 1	–
	Maximum entropy	5	5.2 [2.8, 7.5]	30 / 4	–
	Full-pair copula	5	3.2 [0.5, 6.0]	24 / 5	–
	Equicorrelated copula	5	-3.5 [-6.4, -0.8]	13 / 5	–
	Maximum disagreement	5	-14.7 [-17.9, -11.7]	3 / 2	–
	Maximum entropy	8	-1.5 [-3.7, 0.4]	17 / 5	–
	Full-pair copula	8	-0.8 [-2.6, 1.0]	15 / 6	–
	Equicorrelated copula	8	-7.0 [-9.4, -4.6]	7 / 5	–
	Maximum disagreement	8	-16.3 [-19.8, -12.9]	4 / 2	–
RewardBench 2 coverage, 165 judges	Accuracy ranking	3	reference	–	30.0
	Maximum entropy	3	90.7 [83.1, 97.0]	–	1.5
	Full-pair copula	3	88.3 [80.6, 95.0]	–	1.5
	Equicorrelated copula	3	61.0 [46.8, 77.1]	–	10.5
	Maximum disagreement	3	66.4 [40.1, 86.3]	–	5.5
	Accuracy ranking	5	reference	–	22.0
	Maximum entropy	5	98.8 [97.7, 99.8]	–	0.0
	Full-pair copula	5	94.9 [89.4, 99.3]	–	0.0
	Equicorrelated copula	5	80.3 [71.8, 89.1]	–	3.0
	Maximum disagreement	5	89.4 [74.9, 97.6]	–	1.0
	Accuracy ranking	8	reference	–	16.0
	Maximum entropy	8	99.3 [97.8, 100.0]	–	0.0
	Full-pair copula	8	100.0 [100.0, 100.0]	–	0.0
	Equicorrelated copula	8	96.5 [90.3, 100.0]	–	0.0
	Maximum disagreement	8	99.4 [98.3, 100.0]	–	0.0

Majority vote, strict prompt accuracy (%)	k=3	k=5	k=7	Gain at k=7 [95% CI]
Top-judge committee (reference)	72.1	72.5	72.0	–
Maximum entropy	70.3	73.9	75.0	+3.0 [+1.3, +4.7]
Independence	70.7	72.1	72.2	+0.1
Random committees	64.7	66.7	67.4	-4.6

A controlled comparison with the complete response matrix. In a second protocol, every selector scores the same candidate set at each k: every group when there are at most 300, otherwise 300 groups drawn with a fixed, outcome-independent seed. Selection uses the fit half and the chosen group is scored on the other half. Mean held-out collective failure, in percentage points, is

15.69, 12.19, and 10.03 for maximum entropy and 15.56, 12.03, and 9.95 for counting collective failure directly from the fit half's complete response matrix, so direct counting minus maximum entropy is -0.12, -0.15, and -0.08 points (from unrounded means), with intervals $[-0.53, 0.45]$, $[-0.96, 0.07]$, and $[-0.84, 0.57]$ when the five releases are resampled. Accuracy ranking gives 17.64, 14.56, and 11.58. Direct counting is therefore marginally better than maximum entropy when both see the same candidates, and both are better than accuracy ranking.

Release-level intervals. Resampling the five releases as clusters instead of the 48 pools gives reductions of 0.5%, 14.9%, and 22.7% for maximum entropy against accuracy ranking, with intervals $[-33.5, 24.9]$, $[-6.3, 32.2]$, and $[0.9, 42.0]$ at $k = 3$, 5, and 8. For the strong-quarter setting the values are -2.2%, 8.8%, and 16.0%, with intervals $[-30.5, 18.5]$, $[-2.5, 20.2]$, and $[7.4, 24.0]$.

H AGENTIC CODING POOLS

Data. SWE-bench Verified results come from the benchmark's public results repository, which holds one folder per leaderboard submission. Of the 182 folders, 135 list their resolved tasks and 40 give a resolved flag for each of the 500 tasks. The other seven publish no per-task results and are not used. The task list is the canonical set of 500, so the 27 tasks no system solves are kept. Submissions that agree on at least 92% of tasks are merged, keeping the earliest. The rule is applied only to submissions resolving at least 2% of tasks, since two runs that resolve almost nothing agree on almost every task. Six sets are merged: Claude Sonnet 4 under five agents, Claude Opus 4.5 under three, two versions of one GPT-5 agent, two SWE-Fixer submissions, Gemini 3 Pro under two agents, and three retrieval baselines built on different models that agree only because they fail almost every task. The last merge is the least justified. It changes little, since those runs resolve 3 to 7% of tasks. The merges remove 11 submissions, leaving $175 - 11 = 164$ systems.

One agent, different models. Forty-one submissions run the same agent, mini-SWE-agent, and differ only in the model. Reconstruction on this subset matches the full pool, where maximum entropy misses by 1.3% at three models and 8.0% at eight against compatible-interval half-widths of 6.6% and 46.6%: on the subset it misses by 1.2% and 8.1% against half-widths of 7.8% and 50.6%, and the full-pair copula by 1.7% and 9.1%. Its error on coverage is 0.4% and 1.5%.

Choosing a group. Table 12 reports group choice on this collection with the steps of Appendix G unchanged, over twenty random splits of the 500 tasks in both directions. On the full pool, maximum entropy is lower than the most accurate systems on 31, 32, and 31 of the 40 split directions at $k = 3$, 5, and 8. On the one-agent subset the reductions are 7.3%, 5.2%, and -1.5% (Table 12). The splits share the same 500 tasks and are not independent samples. Without the merge rule, on all 175 submissions, the maximum-entropy reductions are 9.2%, 13.0%, and 8.6% (lower on 32, 33, and 26 of the 40 directions), the full-pair copula gives -0.4%, -4.0%, and -6.4%, the equicorrelated copula -9.8%, -13.6%, and -18.7%, and maximum disagreement -11.3%, -8.7%, and -14.2%, so the merge rule moves the maximum-entropy reduction by at most 2.4 points and changes no ordering. A metadata-only merge was not run, because the run names carry no separate agent field.

SWE-bench Pro. The public set (Deng et al., 2025) has 731 tasks. Scale's evaluation outputs for its leaderboard runs are in a public bucket, and we mark a task resolved when every test it requires passes, the rule of Scale's own evaluation code. The resulting resolve rates for Claude Sonnet 4.5, Claude Haiku 4.5, Claude Sonnet 4 and GPT-5 fall within the published confidence intervals, and GLM-4.5 has no published entry to compare against, since the leaderboard lists GLM-4.6 only. The pool holds the five dated leaderboard runs: Claude Sonnet 4.5, Claude Haiku 4.5, Claude Sonnet 4, GLM-4.5, and GPT-5. One run lacks outputs for 178 tasks and the other four for 3 to 10 each, so the pool keeps the 546 tasks with an output from all five runs. With five models there are 10 groups of three, 5 of four and 1 of five. Maximum entropy misses β by 1.6%, 4.0% and 6.7% against compatible-interval half-widths of 7.9%, 17.5% and 28.2%, and the full-pair copula by 1.4%, 3.0% and 4.3%. Five models leave no real choice of group, so selection is not tested on this pool.

I SCOPE LIMITS

A pairwise table measured on one benchmark does not predict collective failure on another. We tested this in a pre-specified design across three separate collections of pools, holding out one task at a time and giving every method the target task's individual failure rates. The collections are the 21 scenarios of the HELM re-release, which share model identifiers across scenarios (177 distinct models, 27 to 91 per scenario, so the HELM MMLU pool from the separate per-instance store is not included), the 10 EmbedLLM tasks (the same 112 models in each), and 4 Open LLM Leaderboard tasks (304 to 346 models each). Groups are 30 random groups per target task, kept only when the target shows at least 20 collective failures and every pair has pooled statistics from at least two source tasks. At eight models, maximum entropy with pairwise statistics pooled from the other tasks misses the held-out task's collective failure rate by 83%, 44%, and 36% on the three collections, a shuffle control that destroys pair identity while preserving the same multiset of inputs by 89%, 46%, and 37%, and the same estimator with the target task's own pairs by 13%, 6%, and 7%. The pooled-minus-shuffle differences are -5.5, -1.8, and -0.8 points, with paired bootstrap intervals over target tasks of $[-14.0, 2.6]$, $[-6.5, 2.3]$, and $[-3.4, 0.9]$, each containing zero. The own-pairs errors are lower than the HELM value of Table 4 because the event floor removes the rare-event groups. What pooled statistics carry is only the overall level of coupling, and even that is better measured on the target task. Every collective-failure reconstruction and coverage-selection result in this paper is therefore within-pool, and the RewardBench 2 majority-vote study is a separate exploratory leave-one-skill-out decision experiment. That study asks only whether the ordering of candidate committees carries to a held-out skill, not whether a rate transfers, which is the weaker requirement stated in Section 5. Reconstruction and held-out prediction differ in whether the same items or separate halves supply the pairwise table and the target.

The reconstruction also needs the pairs to be well measured. We scored each model on an independent random fraction f of a benchmark's items, the observation fraction, as in pooled logs where each model was scored on its own random subset of items, so that a pair is measured on about f^2 of the items. This setting describes logs that are already incomplete, not a recommended evaluation design. Groups are 40 random groups per pool at three, five, and eight models across the 48 pools, drawn for this experiment and so not the groups behind Table 4. On the five-model groups, at observation fractions $f = 1$ and $f = 0.5$ maximum entropy is the most accurate completion, at 7.7% and 24.2% against 11.1% and 27.0% for the full-pair Gaussian copula and 18.8% and 27.4% for the equicorrelated copula. At $f = 0.3$ the pairs are too noisy to separate the completions and all three read near 31%. Direct counting is possible only on the items scored by all members, a median of 30 at $f = 0.5$ and 2 at $f = 0.3$ against 236 and 77 shared by a pair. On the groups where it is defined (99% and 75% of the five-model groups, 82% and 11% of the eight-model groups), it misses by 65.6% and 109.9% at five models against 24.2% and 29.0% for maximum entropy on the same groups, and by 129.2% against 31.7% at eight models and $f = 0.5$. It reads zero on 25% and 37% of the five-model groups. The same number of evaluations spent on one common random subset removes this advantage (Table 13): at $f = 0.5$, direct counting on the common subset misses by 9.9%, 13.6%, and 17.3% at three, five, and eight models, maximum entropy on it by 10.0%, 15.3%, and 23.4%, and the full-pair copula by 15.6%, 18.0%, and 28.2%, all below the independent-subset values. The advantage of the pairwise table is therefore a property of logs that are already incomplete in this way.

Independently sampled moments need not be jointly consistent. After each pair is projected into its Fréchet bounds, the linear program of Section 2 finds no joint distribution for 7.6%, 28.9%, and 49.2% of the tables at three, five, and eight models at $f = 0.5$, and for 14.9%, 54.2%, and 80.0% at $f = 0.3$. On every such table the bounded fitter stops without matching the moments (median gradient norm 0.02 to 0.06) and returns an approximate fit whose moments remain mismatched, which is what the numbers above use. Projecting the moments to the nearest feasible vector in L_1 (median shift 0.02 to 0.05) and refitting changes the mean error on those tables by at most five points in either direction (Table 13), so the results are reported without projection.

Table 13: **Sparse logs: independent subsets against one common subset of the same size.** Relative error in percent of the all-item collective failure rate, averaged within pools (pools with at least ten usable groups) and then across pools, on 40 random groups per pool and size. Design A: each model scored on its own random fraction f of the items. Direct counting uses the items scored by every member and is reported on the groups where such items exist (defined, in percent of groups), with maximum entropy on the same groups for the paired comparison. Design B: every model scored on one common random subset of $\lfloor fn \rceil$ items. Infeasible: share of design-A tables, after projecting each pair into its Fréchet bounds, that no joint distribution matches; on every such table the bounded fitter stops without matching the moments. Before / after: maximum entropy on those tables as fitted and after projection to the nearest feasible moment vector in L_1.

		design A, all groups			design A, paired			design B, common subset			infeasible tables		
k	f	max. ent.	copula	direct	defined	max. ent.	direct	direct	max. ent.	copula	share	before	after
3	0.5	16.5	20.2	24.3	100%	16.5	25.1	9.9	10.0	15.6	8%	27.7	23.2
3	0.3	25.2	25.9	47.9	99%	25.2	48.1	16.1	16.3	19.9	15%	28.2	31.7
5	0.5	24.2	27.0	61.1	99%	24.2	65.6	13.6	15.3	18.0	29%	31.2	32.3
5	0.3	31.6	31.2	109.5	75%	29.0	109.9	21.7	24.4	22.7	54%	40.5	40.8
8	0.5	34.1	35.5	141.5	82%	31.7	129.2	17.3	23.4	28.2	49%	43.3	43.4
8	0.3	46.2	47.3	104.2	11%	12.8	114.7	27.3	33.9	34.0	80%	51.1	52.3