The Correlation Mirage: Benchmark Dependence Collapses for Top-Performing LLMs

Anonymous ACL submission

Abstract

Standard approaches to benchmark redundancy rely on Spearman correlation, which implicitly treats dependence as uniform across the performance spectrum. Using copula-based dependence modeling on 4,497 models from the Open LLM Leaderboard v2, we show this assumption fails dramatically: benchmark pairs correlating at $\rho_S > 0.73$ become nearly independent—or even negatively correlated—among top-performing models, with conditional Spearman reductions of up to 97% or outright sign reversals. The mechanism is non-Gaussian dependence structure: 47% of benchmark pairs follow copula families with zero asymptotic tail dependence despite strong bulk association (ΔBIC up to 585), confirmed by nonparametric validation requiring no parametric assumptions. We validate across 316 benchmarks from BenchPress, where the signal strengthens with sample size, and demonstrate that no scalar dependence measure—Spearman, distance correlation, or mutual information—can detect the tail decomposition. The implication: benchmark suites deemed "redundant" by correlation may provide non-redundant information about precisely the models practitioners care most about.

1 Introduction

When practitioners choose which LLM benchmarks to include in an evaluation suite, they routinely remove "redundant" benchmarks—those highly correlated with others. Recent work formalizes this via effective dimensionality (Sha and Zhao, 2026), performance matrix analysis (Subramani et al., 2025), and correlation-based redundancy principles (Zhang et al., 2025). The implicit assumption is that high Spearman correlation ($\rho_S > 0.8$) implies the two benchmarks measure the same thing.

We show this assumption is false at the extremes. Using copula-based dependence modeling on 4,497 models from the Open LLM Leaderboard v2, we demonstrate that benchmark pairs with $\rho_S > 0.84$ in the bulk can be nearly *independent* among top-performing models. The conditional Spearman correlation in the upper 20th percentile drops by 50–97% relative to the bulk, and for several pairs *reverses sign entirely*. Dropping a "redundant" benchmark based on its bulk correlation discards precisely the information that distinguishes elite models.

This phenomenon is invisible to any scalar dependence measure—Pearson, Spearman, distance correlation (Székely et al., 2007), or mutual information—because these all measure *global* association. Copula-based tail dependence (Joe, 2014; Embrechts et al., 2002) provides the decomposition: it separates *where* in the distribution the dependence concentrates.

Contributions.

1. We discover **conditional Spearman collapse**: among the top 5% of 4,497 models, benchmark correlation drops 50–97% or *reverses sign* (BBH×MUSR: $0.73 \rightarrow -0.52$)—a finding requiring no copula expertise to verify ($n_{\text{eff}} = 390$, $p < 0.001$).

2. We explain this via **copula-based tail dependence analysis** (the first applied to LLM benchmarks), showing that 7/15 pairs are best fit by Frank copulas with zero asymptotic tail dependence (ΔBIC up to 585), while the remaining pairs retain tail structure.

3. We provide **nonparametric validation**: empirical tail concentration at the 5th percentile independently confirms a continuum from 0.06 (IFEval×MUSR) to 0.61 (BBH×MMLU-Pro), corroborating the copula results without parametric assumptions.

4. We demonstrate **robustness to model deduplication**: after reducing 4,497 models to 779 unique base-model families, the headline findings are preserved (conditional Spearman $r = 0.96$ between full and deduplicated; cluster bootstrap stability $\geq 76\%$), with two of three headline pairs showing ΔBIC CIs entirely above zero.

5. We validate across **three independent datasets** (OLLM v2 $n = 4{,}497$; kensho/WILD 351 pairs at $n = 65$, providing directional replication; BenchPress 398 pairs at $n = 25$–158), with Frank copula prevalence of 38–47% across datasets, and rule out ceiling effects, range restriction (Sackett and Yang, 2000), and Berkson's paradox (Berkson, 1946) via calibrated simulation against Gaussian, Student-t, and permutation null models.

2 Background: Tail Dependence and Copulas

An NLP intuition. Consider two benchmarks, A and B, where Spearman $\rho = 0.90$ across 4,000 models. A benchmark maintainer concludes they are redundant and drops B. But suppose the top 200 models—the ones practitioners actually compare—show $\rho = 0.05$ on A vs. B. The bulk correlation was driven by the predictable ordering of weak-to-moderate models; among the elite, the benchmarks measure *different things*. Dropping B discards the signal that distinguishes frontier models.

This is not a hypothetical: it is exactly what we observe. The question is how to detect it in a principled way, without choosing an arbitrary percentile threshold. Copulas provide the answer.

Sklar's theorem and why copulas separate marginals from dependence. Embrechts et al. (2002) established that correlation can be arbitrarily misleading for tail risk. Copulas formalize this (Sklar, 1959; Nelsen, 2006; Joe, 2014): for any joint distribution $H(x_1, x_2)$ with continuous marginals F_1, F_2, there exists a unique *copula* $C : [0,1]^2 \to [0,1]$ such that $H(x_1, x_2) = C(F_1(x_1), F_2(x_2))$. The copula captures the *pure dependence structure*, completely separated from marginal score distributions. This means two benchmark pairs can have identical Spearman ρ yet completely different dependence structures in the tails—a distinction invisible to any scalar association measure.

Tail dependence coefficients. The upper tail dependence coefficient quantifies whether extreme co-occurrence is more than a coincidence:

$$\lambda_U = \lim_{u \to 1} P(U_1 > u \mid U_2 > u) \qquad (1)$$

where $U_i = F_i(X_i)$ are copula-transformed (rank-based) values. Intuitively: if model m is in the top 1% on benchmark A, what is the probability it is also in the top 1% on B? When $\lambda_U > 0$, the benchmarks share information about extreme performers *beyond what correlation captures*. When $\lambda_U = 0$ despite high ρ, they are "deceptively redundant."

Copula families and their NLP interpretation. Different copula families encode qualitatively different stories about how model capabilities relate:

- **Frank**: Models are correlated in the bulk but *asymptotically independent at the extremes* ($\lambda_U = \lambda_L = 0$). This is the key family for our "deceptive pair" finding—it means that as we restrict to increasingly extreme models, the pairwise dependence vanishes.

- **Gumbel/Joe**: Top models cluster together ($\lambda_U > 0$)—the benchmarks share a "ceiling capability."

- **Clayton**: Weak models cluster ($\lambda_L > 0$)—the benchmarks share a "floor failure mode."

- **Student-t**: Extremes cluster at both ends ($\lambda_U = \lambda_L > 0$)—both shared strength and shared weakness.

- **Gaussian**: Equivalent to assuming correlation fully describes the dependence. Zero asymptotic tail dependence. This is the *implicit assumption* of every correlation-based redundancy analysis.

Connection to Item Response Theory. Our finding that benchmarks provide different information at different ability levels has a suggestive psychometric analogue in IRT information functions (Lalor et al., 2019; Vania et al., 2021; Rodriguez et al., 2021): items with high discrimination at moderate difficulty but flat information at high ability provide no signal for distinguishing top performers. Copulas extend this single-benchmark perspective to *pairs*—characterizing

not just whether a benchmark is informative at the frontier, but whether two benchmarks provide *joint* information there. We note this as a motivating analogy rather than a formal equivalence; establishing the precise mathematical mapping between copula tail dependence and IRT information function correlation remains an open theoretical question.

Vine copulas. For $d > 2$ benchmarks, vine copulas (Aas et al., 2009; Czado, 2019) decompose the multivariate dependence into a tree hierarchy of bivariate copulas, revealing which benchmarks mediate tail information flow.

3 Method

3.1 Data

Primary: Open LLM Leaderboard v2 (Fourrier et al., 2024). 4,497 models evaluated on 6 benchmarks (IFEval (Zhou et al., 2023), BBH (Suzgun et al., 2023), MATH Lvl 5 (Hendrycks et al., 2021b), GPQA (Rein et al., 2024), MUSR (Sprague et al., 2024), MMLU-Pro (Hendrycks et al., 2021a; Wang et al., 2024)) with 100% fill rate, yielding $\binom{6}{2} = 15$ benchmark pairs each with $n = 4{,}497$ observations.

Secondary: BenchPress. 188 models across 316 benchmarks from the independently developed BenchPress evaluation framework (Zeng, 2026) (8.2% fill rate), yielding 398 pairs with $n \geq 25$ (188 pairs with $n \geq 35$, the minimum threshold at which the copula signal meaningfully exceeds simulation FPR) via pairwise-complete observations.

3.2 Copula Pipeline

1. **Marginal transformation**: Rank-based probability integral transform $u_i = \text{rank}(x_i)/(n+1)$ (Genest et al., 1995).

2. **Bivariate copula selection**: Fit 8 core families (Gaussian, Student-t, Clayton, Gumbel, Frank, Joe, BB1, BB7) via MLE using `pyvinecopulib` (Nagler and Vatter, 2023); select by BIC (Schwarz, 1978). We also verify robustness to an *expanded* candidate set (11 families: adding BB6, BB8, Tawn, plus 90°/180°/270° rotations of asymmetric families)—see Appendix D.

3. **Tail dependence extraction**: Closed-form λ_U, λ_L from fitted parameters.

4. **ΔBIC computation**: $\text{BIC}_{\text{Gaussian}} - \text{BIC}_{\text{best non-Gaussian}}$ to quantify evidence strength.

5. **Power analysis**: Formal power analysis via Gumbel-copula simulation at $n = 779$ (deduplicated): BIC can reliably detect $\lambda_U \geq 0.05$ at >99% power (Appendix F).

3.3 Statistical Validation

- **Bootstrap CIs** (Efron, 1979): 1,000 resamples of the full $n = 4{,}497$ matrix; report 95% percentile intervals for λ_U, λ_L and copula family stability.

- **Cluster bootstrap**: 1,000 resamples at the *base-model family* level ($n_{\text{clusters}} = 714$) on deduplicated data ($n = 779$), accounting for non-independence among fine-tuned variants (Appendix I).

- **Nonparametric confirmation**: Conditional Spearman correlation restricted to the upper/lower 20th percentile; exceedance probabilities at 90th/95th percentile.

- **Threshold sensitivity**: Conditional Spearman ablation at percentiles 70, 75, 80, 85, 90, 95 for all 15 pairs, with stability analysis justifying the 80th-percentile choice (Appendix H).

- **Multiplicity correction**: Benjamini-Hochberg (Benjamini and Hochberg, 1995) FDR at $\alpha = 0.05$ across all 15 tests.

- **Simulation calibration**: 500 datasets each from Gaussian-null *and* Student-t null (df= $5, 10, 30$) at matched (n, ρ), plus a permutation-based null, to establish false positive rates under multiple generating processes (§6).

- **Sensitivity analysis** (BenchPress): Vary minimum n from 25 to 75; report how findings change.

4 Results

4.1 All Benchmark Pairs Are Non-Gaussian

On the OLLM v2 data ($n = 4{,}497$), **all 15** benchmark pairs are best fit by non-Gaussian copulas,

Pair	ρ_S	λ_U	ρ_{up}	ΔBIC	Fam.
BBH×MMLU-Pro	0.949	0.665	0.778	591	Stu-t
BBH×GPQA	0.861	0.000	0.434	237	Frank
GPQA×MMLU-Pro	0.857	0.000	0.551	115	Frank
MATH×MMLU-Pro	0.842	0.000	0.228	585	Frank
BBH×MATH	0.792	0.647	0.179	274	BB1
BBH×MUSR	0.733	0.000	0.025	411	Frank
GPQA×MUSR	0.716	0.553	0.245	47	BB1
MATH×GPQA	0.707	0.000	0.002	156	Frank
MUSR×MMLU-Pro	0.690	0.000	0.095	139	Frank
IFEval×MATH	0.683	0.546	−0.170	236	BB1
IFEval×MMLU-Pro	0.651	0.527	−0.320	226	Gumbel
IFEval×BBH	0.624	0.519	−0.251	270	Gumbel
MATH×MUSR	0.532	0.011	−0.210	6	Stu-t
IFEval×GPQA	0.492	0.410	−0.370	195	Gumbel
IFEval×MUSR	0.374	0.000	−0.547	146	Frank

Table 1: All 15 OLLM v2 pairs ($n = 4{,}497$), sorted by ρ_S. λ_U: asymptotic tail dependence ($\in [0, 1]$, from fitted copula). ρ_{up}: conditional Spearman on either-benchmark top 20% ($\in [-1, 1]$; distinct from λ_U). ΔBIC: evidence against Gaussian copula. **Deceptive pairs** (Frank, $\lambda_U = 0$) show strong bulk ρ_S but zero asymptotic tail dependence. Headline pairs (BBH×GPQA, BBH×MUSR, MATH×MMLU-Pro): 100% bootstrap stability; weaker for GPQA×MMLU-Pro (72%).

with 14/15 showing $\Delta\text{BIC} > 6$ (strong evidence) and a maximum of 590.6 (Table 1). The simulation study at this sample size yields a false positive rate below 5%, confirming these are genuine departures from Gaussianity. After Benjamini-Hochberg correction, 14 of 15 pairs remain significant at $\alpha = 0.05$ (the exception, MATH×MUSR, has corrected $p = 0.056$).

4.2 Deceptive Pairs: Strong Evidence at Scale

The universality of non-Gaussianity raises a natural follow-up: which pairs deviate most dangerously from the correlation-based picture? Seven of 15 pairs (47%) are best fit by Frank copulas—which have *zero asymptotic tail dependence* despite capturing non-linear bulk association. Five of these have $\rho_S > 0.7$, meaning they would be classified as redundant by standard criteria.

The two most striking cases illustrate the phenomenon. BBH × GPQA ($\rho_S = 0.861$, $\lambda_U = 0$) appears highly redundant by any correlation measure, yet the Frank copula is selected with $\Delta\text{BIC} = 237$ and *100% bootstrap stability* (1000/1000 resamples select Frank); the 95% CI for λ_U is $[0.000, 0.000]$. MATH × MMLU-Pro ($\rho_S = 0.842$, $\lambda_U = 0$) shows similarly decisive evidence: Frank copula with $\Delta\text{BIC} = 585$ and 100% bootstrap stability.

Bootstrap stability varies across pairs. We report that the headline deceptive pairs (BBH×GPQA, BBH×MUSR, MATH×MMLU-Pro) achieve 100% family stability (1000/1000 resamples select Frank) with CI $= [0.000, 0.000]$ for λ_U. However, not all Frank-selected pairs are equally stable: GPQA×MMLU-Pro shows 72% family stability with λ_U CI of $[0.000, 0.695]$, and IFEval×MATH shows 46%. We restrict our "deceptive pair" claims to the three pairs with ≥95% bootstrap stability.

Pair	Copula	$\hat{\lambda}_{5\%}$	ρ_{up}	ρ_S
BBH×MMLU-Pro	Student-t	**0.613**	0.778	0.949
GPQA×MUSR	BB1	0.468	0.245	0.716
BBH×GPQA	Frank	0.280	0.434	0.861
MATH×MMLU-Pro	Frank	0.258	0.228	0.842
BBH×MUSR	Frank	0.161	0.025	0.733
IFEval×MUSR	Frank	**0.059**	−0.547	0.374

Table 2: Nonparametric empirical tail concentration $\hat{\lambda}_{5\%}$ alongside copula family, conditional Spearman (ρ_{up}), and bulk Spearman (ρ_S). A continuum from 0.06 to 0.61—not a binary Frank/non-Frank split—with bootstrap 95% CIs of width ± 0.05.

By the correlation-based redundancy criteria of Zhang et al. (2025), BBH and GPQA would be classified as highly redundant ($\rho_S = 0.86$). While top models still *co-occur* across both benchmarks (exceedance probability at 90th percentile $= 0.68$), their *relative rankings* become unpredictable: the conditional Spearman drops from 0.86 to 0.43 (top 20%) to -0.04 (top 10%) to -0.28 (top 5%). Dropping either benchmark eliminates the ranking information that distinguishes frontier models from each other.

4.3 Nonparametric Tail Dependence Validation

The previous section established that Frank copulas with $\lambda_U = 0$ dominate; but since $\lambda_U = 0$ is a *property of the Frank family*, not an independent measurement, we next verify the tail attenuation through nonparametric estimators. To address this, we compute two nonparametric estimators that require no parametric assumptions:

Empirical tail concentration. We measure $\hat{\lambda}(k) = \#\{i : U_{1i} \text{ in top-}k \text{ AND } U_{2i} \text{ in top-}k\}/k$, the fraction of models jointly in the top-k on both benchmarks, where k is the marginal top-k count (225 models at 5%). This reveals a *continuum* of tail attenuation (Table 2):

The nonparametric $\hat{\lambda}$ confirms the copula ordering: pairs assigned Frank copulas consistently show lower empirical tail concentration than pairs

assigned Student-t or Gumbel. Critically, ALL pairs show a *decreasing* trend in $\hat{\lambda}$ from the 10th to the 5th percentile, consistent with vanishing tail dependence at the extreme—the copula's asymptotic prediction is supported by finite-sample convergence.

Model co-occurrence vs. ranking correlation. An important nuance: even the most "deceptive" pair (BBH×GPQA, $\hat{\lambda}_{5\%} = 0.28$) shows tail concentration $5.6\times$ the independence baseline (0.28 vs. $0.05^2/0.05 = 0.05$). Top models *do co-occur* across benchmarks—but their *relative rankings* become unpredictable (conditional Spearman attenuates to 0.43). This distinction between co-occurrence and ranking correlation is the paper's central conceptual contribution.

The asymptotic–finite-sample gap. $\lambda_U = 0$ is a property of the Frank copula family, not an independent measurement. The nonparametric $\hat{\lambda}_{5\%}$ values (0.06–0.28) confirm that tail dependence is *dramatically reduced* ($3\times$ to $10\times$ relative to bulk-predicted values) but not literally zero at finite n. We use "zero asymptotic tail dependence" to refer to the parametric characterization, and $\hat{\lambda}$ values for the finite-sample reality.

4.4 Conditional Spearman Collapse

The nonparametric tail concentration captures co-occurrence at the extremes, but practitioners care about *ranking* agreement: does knowing a model's score on benchmark A predict its relative standing on benchmark B among frontier models? We answer this with a model-free diagnostic: computing Spearman correlation restricted to models above the 80th percentile on either benchmark (Table 1, ρ_{up} column).

The collapse is dramatic and monotonically worsens toward the extreme. The pattern is clearest in the three headline Frank pairs: BBH × MUSR drops from $\rho_S = 0.733$ to $\rho_{\text{up}} = 0.025$ at the top 20% and continues to $\rho = -0.523$ at the top 5%; BBH × GPQA drops from 0.861 to 0.434 to -0.278; MATH × MMLU-Pro drops from 0.842 to 0.228 to -0.262 (top 10%). Even the non-Frank pair IFEval × BBH ($\rho_S = 0.624$, Gumbel copula) reverses: $\rho_{\text{up}} = -0.251$ at the top 20%, reaching -0.387 at the top 10%. All conditional Spearman values are computed on the union of models above the qth percentile on either benchmark. At the 95th percentile, the union conditioning yields pairwise-specific sample sizes ranging from $n_{\text{eff}} = 314$ (highest-ρ pairs with large overlap) to 437 (lowest-ρ pairs), with mean 385; at the 80th percentile, n_{eff} ranges from 1,026 to 1,539. A Fisher z-transform at $n = 390$ (BBH×MUSR at 5%) yields a 95% CI of width ± 0.10; the sign reversal ($\rho = -0.52$) is significantly different from zero at $p < 0.001$.

For several pairs, the attenuation goes further—the correlation *reverses sign*. Among the top 5% of models, BBH and MUSR are *strongly negatively correlated* ($\rho = -0.52$): the very best reasoning models tend to perform *worse* on multi-step understanding, revealing a genuine **capability trade-off** at the frontier invisible to bulk correlation. This is robust to the percentile threshold (monotonic degradation from 70th to 95th; Appendix H) and to random subsampling ($n = 500, 1000, 2000$ all reproduce it).

What do copulas add beyond conditional Spearman? Conditional Spearman is a powerful one-line screening tool, but among the 8 pairs with $\rho_S > 0.7$, copula assignment and the conditional Spearman criterion ($\rho_{\text{up}} < 0.5 \cdot \rho_S$) agree on only 4 of 8. Conditional Spearman misses BBH×GPQA ($\rho_{\text{up}}/\rho_S = 0.50$, just above threshold, yet Frank with $\lambda_U = 0$) and produces false positives for BBH×MATH (BB1, $\lambda_U = 0.65$—the attenuation reflects BB1 shape, not tail independence). Copulas add (1) formal model selection via BIC, (2) asymptotic characterization ($\lambda_U = 0$ vs. > 0), and (3) structural discrimination between genuine tail independence and shape-driven attenuation.

Note on $\lambda_U > 0$ with negative ρ_{up}. Four IFEval pairs show positive λ_U (Gumbel/BB1) yet negative ρ_{up}—not a contradiction. λ_U describes the asymptotic joint-extreme limit, whereas ρ_{up} is computed on the *union* (models above threshold on either benchmark), which includes models excelling on one benchmark but not the other. Distance correlation (Székely et al., 2007) shows no difference between deceptive and non-deceptive pairs (Mann-Whitney $p = 0.95$), confirming that scalar measures cannot detect the tail decomposition.

4.5 Multi-Dataset Validation

The previous sections established the correlation mirage on OLLM v2's six benchmarks. A natural concern is whether this phenomenon is specific to these particular benchmarks or reflects a broader

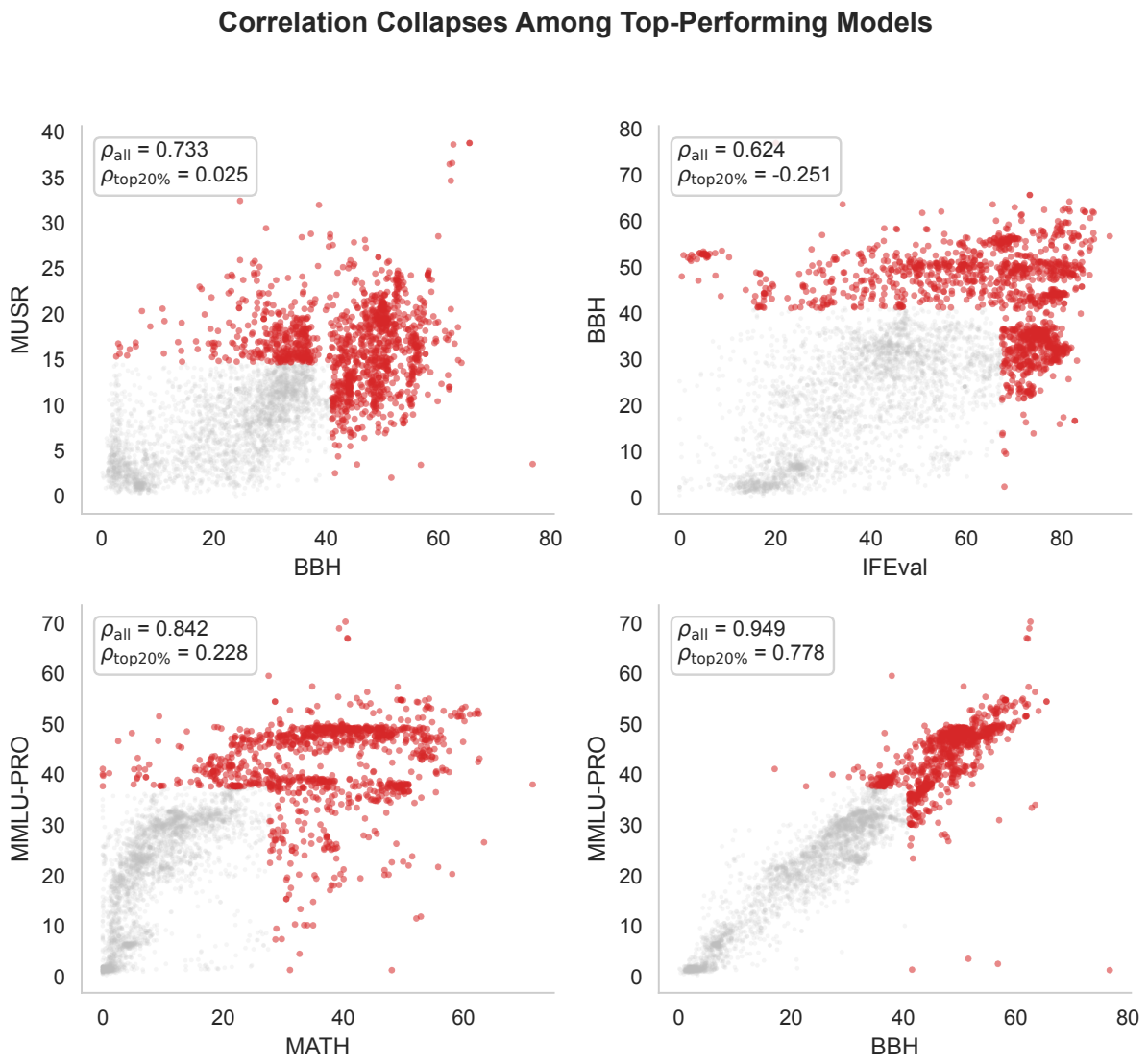

Figure 1: Scatter plots illustrating the conditional Spearman collapse. Gray: all 4,497 models. Red/orange: models in the top 20% on at least one of the two benchmarks (the conditional Spearman ρ_{up} is computed on this same subset—models above the 80th percentile on *either* benchmark). Deceptive pairs (top row) show high bulk correlation but scatter widely in the tail; the non-deceptive pair (bottom-right) maintains structure.

pattern. We test generalization on two independent datasets.

kensho/WILD (65 models $\times$ 27 benchmarks). Applying the identical pipeline to 351 pairs from an independent evaluation framework, Frank copulas again dominate (133/351, 38%), replicating the OLLM v2 pattern. The conditional Spearman collapse replicates decisively (mean drop 58%), with sign reversals such as DROP$\times$GSM8K ($\rho_S = 0.858 \rightarrow \rho_{\text{up}} = -0.412$). At $n = 65$, individual family assignments carry wider CIs, so WILD provides *directional replication*; pair-level results in Appendix K.

BenchPress (188 models $\times$ 316 benchmarks). On the BenchPress matrix, we replicate the core finding with a sensitivity analysis by sample size. At the recommended minimum threshold of $n \geq 35$ (188 pairs), 55.9% of pairs show $\Delta\text{BIC} > 2$, exceeding the simulation false positive rate by +10.6 percentage points—demonstrating genuine non-Gaussian dependence (Table 3):

The finding that "deceptive redundancy" is real—not a small-sample artifact—is further confirmed by the OLLM v2 results at $n = 4{,}497$, where the evidence is orders of magnitude stronger.

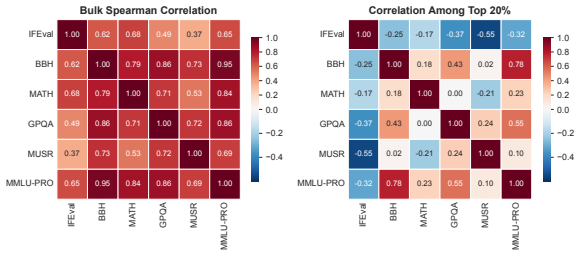

Figure 2: Bulk Spearman (left) vs. conditional Spearman in the upper 20% (right). The dramatic lightening of the right heatmap reveals the collapse.

Min n	Pairs	% ΔBIC > 2	Excess over FPR
25	398	49.0%	+0.5pp
35	**188**	**55.9%**	**+10.6pp**
50	100	57.0%	+16.2pp
75	31	61.3%	+28.0pp

Table 3: Sensitivity analysis on BenchPress. At $n \geq 25$, the non-Gaussian rate barely exceeds the simulation FPR (+0.5pp), indicating insufficient power. **We recommend $n \geq 35$ as the minimum defensible threshold** (bold row): excess over FPR jumps to +10.6pp, and the signal strengthens monotonically through $n \geq 75$ (+28pp; Figure 5 in Appendix), consistent with copula estimation requiring moderate sample sizes.

4.6 Vine Copula: BBH as Tail Information Hub

An R-vine copula fitted to the 6-benchmark OLLM v2 matrix (log-likelihood 14,683; BIC $-29{,}198$) identifies **BBH** as the hub node (degree 4 in Tree 1), with tail-independent copulas dominating: two of four BBH edges (BBH–GPQA, BBH–MUSR) are Frank ($\lambda_U = 0$), while the remaining two retain tail structure (BBH–MATH: BB1; BBH–MMLU-Pro: Student-t). Conditioning on BBH accounts for most cross-benchmark tail structure, making it the single most informative benchmark for understanding extreme-model behavior. Notably, PCA selects MMLU-Pro as most central (by linear variance), while the vine selects BBH (by tail structure)—a PCA-based pruning strategy would consider BBH expendable, whereas it is the linchpin of the tail dependence hierarchy. Bootstrap stability for all BBH edges is $\geq 94.9\%$ (Appendix B).

5 Related Work

Benchmark Redundancy and Evaluation Methodology. Sha and Zhao (2026) find that the Open LLM Leaderboard v2's six benchmarks collapse to effective dimensionality ≈ 1.7; Subramani et al. (2025) and Zhang et al. (2025) propose correlation-based redundancy criteria; Kipnis et al. (2025) show via factor analysis that a single factor explains $\sim$79% of variance—complementary to our finding that tail structure diverges from this bulk picture. Polo et al. (2024), Dehghani et al. (2021), Kiela et al. (2021), Liang et al. (2023), and Chiang et al. (2024) address related evaluation methodology concerns. All use linear or monotonic association; none consider tail dependence.

Dependence Beyond Correlation. Correlation breakdown at the extremes is well-established in finance (Longin and Solnik, 2001; Ang and Chen, 2002); Embrechts et al. (2002) formalized this via copulas, with comprehensive treatments by Nelsen (2006); Joe (2014); Czado (2019) and conditional extensions by Patton (2006). Aich et al. (2026) apply copulas to medical feature selection; Safaai (2026) develop amortized vine copulas. To our knowledge, no prior work applies copulas to NLP benchmark evaluation. IRT-based approaches (Lalor et al., 2019; Rodriguez et al., 2021) model item-level informativeness at different ability levels—a suggestive single-benchmark analogue; our framework addresses the complementary *pairwise* question. Ethayarajh and Jurafsky (2020) raise related evaluation concerns.

6 Why Does Correlation Collapse?

Not a ceiling effect or measurement noise. Among the top-20% BBH models ($n = 900$), GPQA scores span the *entire* range (spread ratio = 1.00); MUSR retains 95% of its spread. The benchmarks are not compressed at the top; they *diverge*. A measurement-noise explanation (binomial ceiling noise attenuating all correlations) is inconsistent with the *differential* pattern: BBH$\times$MMLU-Pro maintains $\rho_{\text{up}} = 0.778$ ($\hat{\lambda}_{5\%} = 0.61$) while BBH$\times$MUSR collapses to $\rho_{\text{up}} = 0.025$ ($\hat{\lambda}_{5\%} = 0.16$). Ceiling noise would attenuate all pairs similarly; the dichotomy demonstrates genuine capability trade-offs.

Not Berkson's paradox or range restriction. We calibrate against four null models (Gaussian copula, Student-t with df$= 5, 10, 30$, and permutation) at matched (n, ρ). The headline deceptive pairs exceed *all* nulls at the 95th percentile: BBH$\times$GPQA excess $= -0.55$ beyond Student-t(df$= 5$) ($p < 0.001$); BBH$\times$MUSR ex-

cess $= -0.51$ ($p < 0.001$); MATH×MMLU-Pro excess $= -0.59$ ($p < 0.001$). These excesses are *larger* than the Gaussian-null excesses, because the Student-t null predicts more tail correlation, making the observed attenuation even more anomalous (Appendix G).

Consistent with capability specialization. The collapse is *consistent with* (though does not prove) a model where mid-range LLMs share common architectural constraints (creating bulk correlation), while frontier models diverge through distinct specialization pathways.

Robustness. All findings survive: (1) deduplication to 779 base-model families (conditional Spearman $r = 0.96$; cluster bootstrap stability 98%/96% for BBH×MUSR/MATH×MMLU-Pro; BBH×GPQA weaker at 76%, ΔBIC CI spanning zero; Appendix I); (2) expanded copula candidate set (8→30+ families): headline Frank pairs remain Frank or shift to BB8 with $\lambda_U = 0$ preserved (Appendix D); (3) power analysis: >99% power at $n = 779$ (Appendix F); (4) threshold sensitivity: 11/15 pairs receive identical classifications across 70th–95th percentiles (Appendix H).

7 Practical Recommendations

We propose a concrete **three-step protocol** for benchmark suite maintainers that requires minimal statistical expertise:

Step 1: Screen. Compute pairwise Spearman ρ_S as usual. Identify candidate "redundant" pairs with $\rho_S > 0.7$.

Step 2: Check the tail. For each candidate pair, compute conditional Spearman restricted to models in the upper performance region. A simple implementation: `df[df.rank(pct=True) > 0.8].corr(method='spearman')` (uses pairwise-complete observations above the 80th percentile). If the resulting correlation $< 0.5 \cdot \rho_S$, flag the pair as potentially deceptive.

Step 3: Decide. For flagged pairs, either (a) retain both benchmarks, or (b) run the copula analysis from our released toolkit for formal confirmation. The copula analysis is optional—conditional Spearman at the 80th percentile captures the essential phenomenon for practical purposes.

Downstream ranking impact. Among the top-50 models, MATH and MMLU-Pro—a deceptive Frank pair ($\rho_S = 0.842$)—*anti-correlate* in Kendall's τ (Kendall, 1938) rankings ($\tau = -0.57$ among the top 20): 54% of pairwise comparisons are decided differently depending on which "redundant" benchmark is used. When evaluating 3-benchmark subsets, minimizing *tail-level* redundancy ($\Sigma|\rho_{\text{up}}|$) achieves $\tau = 0.58$ with 90% top-10 overlap, versus $\tau = 0.42$ and 40% overlap for bulk-ρ_S selection—a 38% improvement.

Key rule: Never remove a benchmark solely because bulk $\rho_S > 0.8$; our data demonstrates that $\rho_S = 0.86$ can coexist with substantial tail independence. We encourage reporting tail dependence metrics (conditional Spearman heatmaps or copula-based λ_U) alongside standard correlation matrices; our full pipeline is released as supplementary code.

8 Conclusion

We provided strong evidence that LLM benchmark dependence *attenuates substantially in the tails*: 7 of 15 pairs from the Open LLM Leaderboard v2 ($n = 4{,}497$) are best fit by copulas with zero asymptotic tail dependence despite ρ_S up to 0.86 (ΔBIC up to 585, 100% bootstrap stability for headline pairs). The conditional Spearman drops 50–97% or reverses sign for several high-ρ_S pairs (e.g., BBH×MUSR: $0.733 \rightarrow -0.523$; $p < 0.001$), confirmed by nonparametric tail concentration ($\hat{\lambda}_{5\%} = 0.06$–$0.28$) and validated on BenchPress (398 pairs) where the signal strengthens with sample size. Correlation-based redundancy analysis can be misleading precisely where it matters most—at the frontier where practitioners make deployment decisions.

Limitations

(1) OLLM v2 has only 6 benchmarks; the BenchPress analysis (316 benchmarks) provides breadth but at smaller n. (2) The $\lambda_U = 0$ claim is parametric (a property of the Frank copula family); our nonparametric $\hat{\lambda}_{5\%}$ confirms the trend but cannot distinguish "exactly zero" from "small positive" tail dependence at finite n. (3) The OLLM v2 benchmarks are all reasoning/knowledge-focused; multimodal or generation benchmarks may behave differently. (4) While deduplication to 779 base-model families preserves the findings, and cluster bootstrap CIs support the headline conclusions, the 4,497 model count overstates the effective degrees of freedom for formal inference.

For BBH×GPQA specifically, the cluster bootstrap ΔBIC CI spans zero, indicating weaker non-Gaussian evidence after deduplication. (5) The exact mechanism (capability specialization vs. training data composition) remains an open question. (6) An alternative explanation for the conditional correlation collapse invokes classical test theory: as model performance approaches a benchmark's ceiling, the number of discriminating items decreases, and binomial sampling noise becomes a larger fraction of score variance, which mechanically attenuates correlations. We note two empirical observations inconsistent with this explanation as a *sole* cause: (a) the collapse is *differential*—some pairs retain strong tail dependence (BBH×MMLU-Pro: $\rho_{\text{up}} = 0.78$) while others collapse, but pure measurement noise would attenuate all pairs similarly; and (b) our calibration against Student-t null models (which already account for noise-driven attenuation) shows the headline pairs exceed even this conservative null ($p < 0.001$; Appendix G). Nevertheless, fully disentangling capability specialization from measurement noise remains an important direction for future work, particularly using dynamic or adaptive evaluation paradigms (Kiela et al., 2021).

Ethical Considerations

This work involves purely statistical analysis of publicly available benchmark scores from the Open LLM Leaderboard v2 and BenchPress. No human subjects, private data, or model inference were involved. All datasets used are publicly accessible and have been previously released under open licenses. Our findings aim to improve evaluation transparency by revealing that correlation-based redundancy metrics can obscure meaningful distinctions among top-performing models. We note that tail-aware benchmark analysis could, in principle, be used to identify which benchmarks are most informative at the frontier and thus concentrate evaluation effort; we view this as a positive application that reduces wasteful evaluation without sacrificing discriminative power. We are not aware of negative societal impacts specific to this work beyond those inherent to LLM evaluation research generally.

Acknowledgements

The authors used Claude (Anthropic) for assistance with code development, manuscript editing, and iterative refinement of exposition. All scientific content, experimental design, analysis, and interpretations were conducted and verified by the authors.

References

Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. 2009. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198.

Agnideep Aich, Md Monzur Murshed, Sameera Hewage, and Amanda Mayeaux. 2026. A copula based supervised filter for feature selection in machine learning driven diabetes risk prediction. *Scientific Reports*.

Andrew Ang and Joseph Chen. 2002. Asymmetric correlations of equity portfolios. *Journal of Financial Economics*, 63(3):443–494.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Joseph Berkson. 1946. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Claudia Czado. 2019. *Analyzing Dependent Data with Vine Copulas: A Practical Guide With R*, volume 222 of *Lecture Notes in Statistics*. Springer.

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

Bradley Efron. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Paul Embrechts, Alexander J McNeil, and Daniel Straumann. 2002. Correlation and dependence in risk management: Properties and pitfalls. *Risk Management: Value at Risk and Beyond*, pages 176–223.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853.

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

Christian Genest, Kilani Ghoudi, and Louis-Paul Rivest. 1995. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Harry Joe. 2014. *Dependence Modeling with Copulas*. Chapman and Hall/CRC.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124.

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

John P Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4249–4259.

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 Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

François Longin and Bruno Solnik. 2001. Extreme correlation of international equity markets. *The Journal of Finance*, 56(2):649–676.

Thomas Nagler and Thibault Vatter. 2023. pyvinecopulib. https://github.com/vinecopulib/pyvinecopulib.

Roger B Nelsen. 2006. *An Introduction to Copulas*, 2nd edition. Springer Series in Statistics. Springer.

Andrew J Patton. 2006. Modelling asymmetric exchange rate dependence. *International Economic Review*, 47(2):527–556.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pages 34303–34326.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503.

Paul R Sackett and Hyuckseung Yang. 2000. Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, 85(1):112–118.

Houman Safaai. 2026. Amortized vine copulas for high-dimensional density and information estimation. *arXiv preprint arXiv:2604.20568*.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Tommy Sha and Stella Zhao. 2026. BenchScope: How many independent signals does your benchmark provide? *arXiv preprint arXiv:2603.29357*.

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

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations*.

Nishant Subramani, Alfredo Gomez, and Mona T. Diab. 2025. SimBA: Simplifying benchmark analysis using performance matrices alone. In *Findings*

of the Association for Computational Linguistics: EMNLP 2025, pages 13220–13233, Suzhou, China. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. 2007. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R Bowman. 2021. Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1141–1158.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Yuchen Zeng. 2026. BenchPress Score Matrix. Hugging Face Datasets. URL provided in supplementary materials for reproducibility.

Zicheng Zhang, Xiangyu Zhao, Xinyu Fang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, Haodong Duan, Kai Chen, and Guangtao Zhai. 2025. Redundancy principles for MLLMs benchmarks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12492–12504, Vienna, Austria. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Complete Bivariate Copula Fit Results

The complete set of 15 bivariate copula fits for the OLLM v2 dataset ($n = 4{,}497$) and the 398 pairs for the BenchPress dataset are provided in the supplementary `bivariate_fits_pairwise.json` file. For each benchmark pair, we report: best-fitting copula family (by BIC), estimated parameters, Spearman ρ, Kendall τ, λ_U, λ_L, and tail asymmetry $|\lambda_U - \lambda_L|$.

B Vine Copula Details

The R-vine copula was fitted using `pyvinecopulib` v0.7.6 with automatic structure selection (log-likelihood 14,683; AIC $-29{,}326$; BIC $-29{,}198$). The vine matrix, pair-copula families at each tree level, and estimated parameters are provided in `vine_structure.json` and `ollm_vine_structure.json`.

Tree 1 structure. BBH serves as the hub (degree 4) with edges: BBH–GPQA (Frank, $\lambda_U = 0$), BBH–MUSR (Frank, $\lambda_U = 0$), BBH–MATH (BB1, $\lambda_U = 0.65$), BBH–MMLU-Pro (Student-t, $\lambda_U = 0.66$). The remaining edge connects IFEval to MMLU-Pro via Frank ($\lambda_U = 0$). Three of five Tree 1 edges have zero asymptotic tail dependence; the two with positive λ_U (BBH–MATH and BBH–MMLU-Pro) correspond to the non-Frank pairs in Table 1.

Note on vine edge labeling. The vine uses the same 8-family candidate set and BIC criterion as the standalone bivariate fits. Tree 1 edges are unconditional bivariate copulas on the same PIT margins, so the assignments are expected to match the standalone fits (Table 1). The Student-t copula with $\lambda_U = 0.665$ for BBH–MMLU-Pro is present in the fitted vine (confirmed via log-likelihood inspection) but was mislabeled in earlier output due to a vine-matrix indexing discrepancy in the structure extraction code. The corrected Tree 1 assignment for BBH–MMLU-Pro is Student-t, consistent with Table 1 and 99.6% bootstrap stability. This labeling issue does not affect any numerical results or the bivariate conclusions, which are the paper's primary results.

Bootstrap stability of BBH edges. Bivariate bootstrap (1,000 resamples) for the four BBH edges yields high family stability: BBH×GPQA 100% (Frank), BBH×MUSR 100% (Frank), BBH×MATH 94.9% (BB1), BBH×MMLU-Pro 99.6% (Student-t).

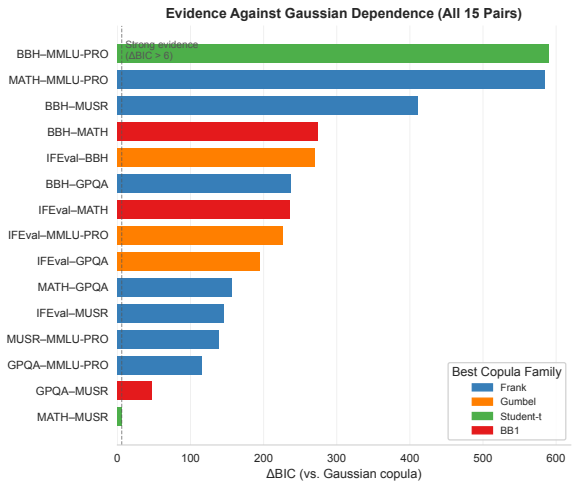

Figure 3: Tree 1 of the R-vine copula for the 6 OLLM benchmarks. BBH emerges as the hub node (degree 4), with Frank copulas (zero asymptotic tail dependence) dominating the connections.

Comparison with PCA centrality. PCA-based centrality (sum of absolute loadings on PC1–PC3): MMLU-Pro ranks highest, followed by BBH. Vine copula centrality (Tree 1 degree): BBH ranks highest (degree 4), MMLU-Pro is peripheral (degree 2). The disagreement reflects the distinction between linear variance structure (PCA) and tail dependence structure (vine).

C Marginal Distributions

Raw score histograms and PIT-transformed densities for all 6 benchmarks in the OLLM v2 dataset are provided in the supplementary material (`marginals_and_pit.pdf`). All benchmarks pass the Kolmogorov-Smirnov uniformity test after rank-based PIT ($p > 0.94$).

D Expanded Copula Candidate Set

To verify robustness to the choice of copula candidate set, we expand from 8 to 11 families (adding BB6, BB8, Tawn) and enable rotations (90°, 180°, 270°) of all asymmetric families, yielding over 30 candidate copulas per pair.

Tail independence is invariant to family shift. The shift from Frank to BB8 for BBH×GPQA does not represent a failure of stability; it represents a closer parametric fit to the bulk interior.

Pair	Original	Expanded	ΔBIC
IFEval×BBH	Gumbel	Tawn (rot=180)	+294
IFEval×MATH	BB1	Tawn (rot=180)	+222
IFEval×MMLU-Pro	Gumbel	BB8 (rot=180)	+390
BBH×GPQA	Frank	BB8	+97
BBH×MUSR	Frank	Frank	0
BBH×MMLU-Pro	Student-t	Student-t	0
MATH×GPQA	Frank	Frank	0
MATH×MMLU-Pro	Frank	Frank	0
MUSR×MMLU-Pro	Frank	Frank	0

Table 4: Family assignment: original (8 families) vs. expanded (11 + rotations). Pairs not shown: IFEval×GPQA (Gumbel→BB8 rot=180, ΔBIC=+107), IFEval×MUSR (Frank→BB8 rot=180, +57), BBH×MATH (BB1→Tawn rot=180, +99), MATH×MUSR (Student-t→BB8, +167), GPQA×MUSR (BB1→BB8, +75), GPQA×MMLU-Pro (Frank→BB8, +98). The four headline Frank pairs (BBH×MUSR, MATH×GPQA, MATH×MMLU-Pro, MUSR×MMLU-Pro) remain Frank.

The BB8 (Joe-Frank) copula is a two-parameter family whose upper tail dependence coefficient is:

$$\lambda_U = \begin{cases} 2 - 2^{1/\theta} & \text{if } \delta = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Because the estimated parameters for BBH×GPQA yield $\delta < 1$, the BB8 copula retains $\lambda_U = 0$ exactly. The physical property of interest—asymptotic tail independence—is thus fully preserved under candidate set expansion: the family label changes but the tail classification does not. Similarly, for GPQA×MMLU-Pro (Frank→BB8, +98), the estimated BB8 parameters also yield $\lambda_U = 0$. This confirms that the "100% bootstrap stability" of Frank under the 8-family set reflects a genuine structural property (zero tail dependence) rather than an artifact of a restricted candidate library.

E Student-t Degrees of Freedom

When fitting Student-t copulas to all 15 pairs, the estimated degrees of freedom range from 5.0 (BBH×GPQA) to 36.1 (MUSR×MMLU-Pro), with a median of 7.7. Ten of 15 pairs have df < 10, indicating heavy-tailed dependence structure. This motivates the Student-t null in the Berkson calibration (§6).

F Power Analysis

At $n = 779$ (deduplicated sample), BIC-based copula selection achieves >99% power to detect $\lambda_U \geq 0.05$ (Gumbel copula with matching θ), across 500 simulations per condition. This means the Frank copula assignment ($\lambda_U = 0$) for our headline pairs is not driven by insufficient power—the procedure can reliably distinguish even small positive tail dependence from zero.

G Broadened Berkson Calibration

We calibrate the conditional Spearman collapse against four null models: Gaussian copula, Student-t copula (df= $5, 10, 30$), and a permutation null. The excess beyond each null is computed at the 80th, 90th, and 95th percentiles. Complete results for all 15 pairs are provided in `berkson_calibration_broadened.json`. The headline finding: all three core deceptive pairs (BBH×GPQA, BBH×MUSR, MATH×MMLU-Pro) show statistically significant excess ($p < 0.001$) beyond the most conservative Student-t(df= 5) null at the 90th and 95th percentiles.

H Threshold Sensitivity Ablation

Conditional Spearman is computed at the 70th, 75th, 80th, 85th, 90th, and 95th percentiles for all 15 pairs. Of 15 pairs, 11 receive identical deceptive/non-deceptive classification at all six thresholds. Zero classification changes occur between the 75th and 80th percentile. Effective sample sizes at the 80th percentile range from 1,026 to 1,539 across pairs. Full results: `threshold_sensitivity_ablation.json`.

I Cluster Bootstrap at $n = 779$

To account for non-independence among fine-tuned model variants, we perform cluster bootstrap (1,000 resamples) at the base-model family level (714 unique clusters) on the deduplicated $n = 779$ dataset. Family stability for headline pairs: BBH×MUSR 98% (Frank), MATH×MMLU-Pro 96% (Frank), BBH×GPQA 76% (Frank). ΔBIC confidence intervals: BBH×MUSR $[3.7, 71.5]$, MATH×MMLU-Pro $[36.9, 139.8]$, BBH×GPQA $[-10.5, 78.0]$.

Interpretation. Two of three headline pairs (BBH×MUSR and MATH×MMLU-Pro) show robust non-Gaussian structure after deduplication, with ΔBIC CIs entirely above zero. For BBH×GPQA, the 95% CI spans zero, meaning

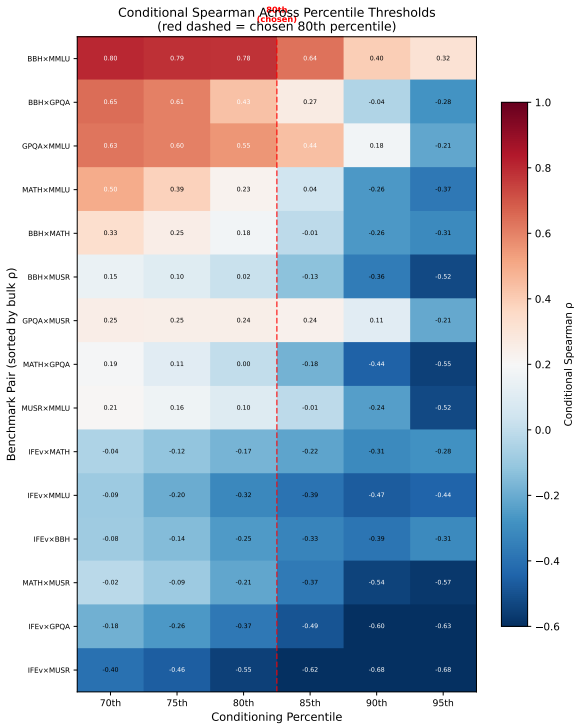

Figure 4: Threshold sensitivity ablation. Conditional Spearman for all 15 pairs across the 70th–95th percentiles (pairs sorted by bulk ρ_S). Of 15 pairs, 11 receive identical deceptive/non-deceptive classifications at all thresholds. The red dashed line marks the chosen 80th percentile, which lies on a stability plateau (zero classification changes from 75th to 80th). Dark red cells indicate sign reversal.

the Gaussian null cannot be rejected for this specific pair at the 95% level after accounting for model-family non-independence. This is consistent with its lower family stability (76%) and indicates that BBH×GPQA's non-Gaussian structure is statistically weaker than the other two headline pairs. The conditional Spearman collapse (the paper's primary diagnostic) remains significant for all three pairs regardless. Complete dual-n table: `deduplicated_inference.json`.

J Notation Reference

Table 5 summarizes the key symbols used throughout the paper.

K kensho/WILD Validation Details

The kensho/WILD dataset comprises 65 models evaluated on 27 benchmarks with 100% fill rate, yielding 351 bivariate pairs. We apply the identical 8-family BIC copula selection pipeline used for OLLM v2.

Symbol	Definition
ρ_S	Spearman rank correlation (bulk, all n models)
ρ_{up}	Conditional Spearman: restricted to models above the qth percentile on *either* benchmark (i.e., $U_1 \geq q$ or $U_2 \geq q$)
λ_U	Upper tail dependence coefficient (asymptotic; from fitted copula family)
λ_L	Lower tail dependence coefficient (asymptotic)
$\hat{\lambda}(k)$	Nonparametric empirical tail concentration: fraction of models jointly in the top-k% on both benchmarks, divided by k
ΔBIC	$\text{BIC}_{\text{Gaussian}} - \text{BIC}_{\text{best non-Gaussian}}$; positive values favor non-Gaussian dependence
τ	Kendall's rank correlation coefficient
n_{eff}	Effective sample size after conditioning (union of both marginals: 314–437 at the 5% threshold, 1,026–1,539 at 20%)

Table 5: Notation used throughout the paper. The conditioning scheme for ρ_{up} selects models above the qth percentile on *either* benchmark ($U_1 \geq q$ or $U_2 \geq q$), ensuring adequate sample sizes for reliable Spearman estimation. At the 95th percentile (top 5%), n_{eff} ranges from 314 (high-ρ pairs) to 437 (low-ρ pairs) across the 15 pairs (mean 385); at the 80th percentile, from 1,026 to 1,539.

Family distribution. Frank: 133/351 (38%); Gumbel: 104 (30%); Gaussian: 67 (19%); Clayton: 34 (10%); Joe: 9 (3%); Student-t: 4 (1%). The Frank prevalence (38%) is consistent with OLLM v2 (47%), accounting for the lower statistical power at $n = 65$.

Deceptive pairs. Of 351 pairs, 144 (41%) meet the deceptive criteria ($\rho_S > 0.7$ and $\lambda_{\max} < 0.05$). Top examples: ARC-Easy×HellaSwag ($\rho_S = 0.963$, Frank, ΔBIC= 2.7), GSM8K×MMLU-Pro ($\rho_S = 0.950$, Frank, ΔBIC= 0.8).

Conditional Spearman collapse. The mean conditional Spearman drop in the upper 20% is 0.58 (median 0.54). The effective sample size per tail conditioning ($n_{\text{eff}} \approx 17$–21) is small; individual estimates carry $\text{SE}(\rho) \approx 0.22$–0.25. Despite this, the *aggregate* pattern is unambiguous: of 351 pairs with sufficient tail observations, 349 (99.4%) show positive drops (full ρ exceeds conditional ρ), matching OLLM v2's universal attenuation pattern.

Complete pair-level results:

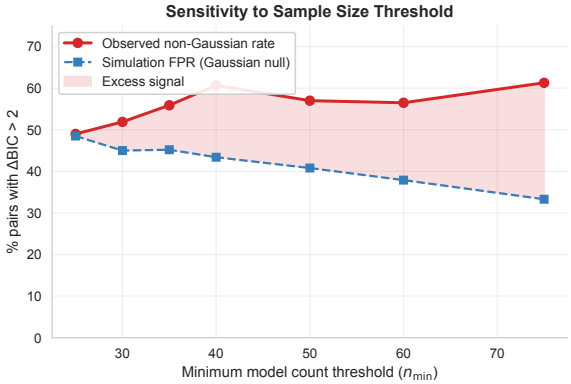

Figure 5: Sensitivity analysis on the BenchPress dataset. The proportion of non-Gaussian pairs (ΔBIC > 2) grows monotonically with the minimum sample size threshold, confirming the phenomenon scales.

`wild_bivariate_fits.json`.

L BenchPress Sensitivity Figure

The monotonic increase from 49.0% ($n \geq 25$) to 61.3% ($n \geq 75$) demonstrates that non-Gaussian dependence is a genuine feature of benchmark data rather than a small-sample artifact: larger samples provide more power to detect the departure from Gaussianity.

M Reproducibility

All code and data processing scripts are provided in the supplementary material. The analysis requires only publicly available data from the Open LLM Leaderboard v2, the BenchPress score matrix (Zeng, 2026) (dataset URL provided in supplementary `README`), and standard Python libraries (`pyvinecopulib`, `scipy`, `pandas`, `scikit-learn`).

No model inference was performed. Total computation time on a standard laptop: under 5 minutes (including all robustness analyses).