INFERRED GENERATIVE-PROCESS DIVERSITY PREDICTS CORRELATED FAILURE ACROSS LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As language models are increasingly deployed in multi-model systems, robust system function depends on constituent models failing independently. Diversity is widely observed to buffer complex adaptive systems against correlated failure, but only when the type of diversity is relevant to the failure mode of concern. Existing assessments of model diversity often focus on semantic differences between outputs. We instead consider *inferred generative-process diversity*, the difference between inferred processes capable of generating observed outputs. We estimate this using Normalised Compression Distance between model output byte strings residualised against a permutation control. Estimated from responses of 38 language models to open-ended chat prompts, the measure predicts which model pairs choose the same wrong answer on ten disjoint benchmark families, beyond semantic similarity and capability, with a negative partial association on every benchmark (mean -0.216, 95% CI $[-0.309, -0.122]$). Semantic distance predicts no reduction in correlated failure once process diversity is held fixed. Five-model ensembles selected to maximise the Vendi score of the measure at fixed semantic distance converge on the same wrong answer less often than capability-matched random ensembles in every semantic band, whereas the reverse selection does not. Because it requires only observed outputs, inferred generative-process diversity provides a practical black-box approach for auditing and constructing model populations where correlated failure creates system-level risk.

1 INTRODUCTION

Language models are increasingly deployed as components of larger systems. Ensembles aggregate their predictions, oversight schemes assign models to critique one another, and multi-agent architectures distribute tasks across many model instances. Such systems are attractive because combining models can improve collective performance and reliability. Complementary capabilities can expand what the system can do, while redundant components can preserve function when one component fails (Kuncheva & Whitaker, 2003; Wood et al., 2023). Neither benefit, however, follows from model count alone. If models share the same weaknesses and reproduce the same errors, an additional vote or reviewer adds nominal redundancy without adding another line of defence. These stakes are now concrete. In the July 2026 Hugging Face incident, nominally isolated agent runs established unauthorised communication and coordinated on harmful activity that no operator had sanctioned, with only rare refusals among the agents involved (Greenblatt et al., 2026; Larcher et al., 2026; OpenAI, 2026). A population of near-identical agents can therefore fail as one, and counting agents, providers, or model families says little about whether it will. What matters is whether the components differ in ways that are relevant to the function and failure modes of the system.

Research on ecological resilience and collective adaptation makes this distinction explicit. Functional diversity can broaden the capabilities available to a collective, whereas response diversity—variation in how components contributing to the same function respond to perturbation—is what enables redundancy to buffer collective function against failure (Holling, 1973; Yachi & Loreau, 1999; Elmqvist et al., 2003). Heterogeneous information and strategies likewise preserve alternative responses as conditions change, whereas homogenisation drives a collective to converge on the same

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.

locally effective behaviour (Galesic et al., 2023; Tieman et al., 2025). This motivates the question, which dimension of model variation is relevant to the failure mode that redundancy aims to mitigate?

Contemporary language models differ in provider, architecture, size, training procedure, and interface, yet these attributes do not ensure behavioural independence. Models from different frontier labs select the same wrong answers more often than expected from their individual error distributions (Goel et al., 2025; Kim et al., 2025). Artificial Hivemind also demonstrates substantial semantic similarity among open-ended responses across model populations (Jiang et al., 2025). Semantic similarity captures whether outputs express similar meanings. It does not determine whether models share the same generative regularities or will make the same error on a separate task. Models can produce semantically similar answers through different processes, or semantically divergent answers despite shared process-level regularities.

We introduce *generative-process diversity* to distinguish these questions. A language model is a stochastic process that maps a prompt and context to a distribution over output sequences, which can be extended to action–observation sequences through addition to an agentic harness. Generative-process diversity concerns differences between processes capable of generating observed behaviour, rather than differences in one selected property of the resulting outputs. As the internal mechanisms of proprietary models are generally unavailable, inferring process-level relations from outputs offers a practical avenue for comparison. Algorithmic Information Theory (AIT) motivates a feature-free estimate through shared description length. We approximate this via Normalised Compression Distance (NCD) of response byte strings residualised against a permutation control, which preserves byte frequencies while destroying order (Cilibrasi & Vitányi, 2005; Vitányi et al., 2009).

We evaluate the measure using a task-transfer design. We focus on language-model input–output behaviour without additional agent scaffolding or tooling as an initial test of generative process diversity. Diversity is estimated from repeated responses by 38 language models to 100 open-ended prompts from the Infinity-Chat taxonomy (Jiang et al., 2025). We then test whether this independently derived population geometry predicts cross-task variation in correlated wrong answers among pairs in the same model population on ten disjoint closed-form benchmark families, controlling for semantic similarity and capability. We then ask whether the pairwise signal survives at the level of five-model ensembles when semantic distance is held fixed by construction. We make three contributions:

1. Define inferred generative-process diversity relative to a system function and estimate it using permutation-control-residualised NCD.
2. Show the permutation control residualisation separates an order-specific component from marginal output composition, that these components have opposing relationships with correlated failure, and that the order-specific component is related to, but non-redundant with, semantic similarity.
3. Demonstrate that inferred generative-process diversity predicts cross-task variation in correlated failure among evaluated model pairs beyond semantic similarity and capability. Under the cross-controlled specification, the association is negative on all ten benchmark families with a cross-benchmark mean of -0.216 and 95% interval of $[-0.309, -0.122]$. At the ensemble level, five-model sets selected for generative-process diversity at a fixed level of semantic distance show lower chance-corrected convergent failure than capability-matched sets averaged over nine benchmarks, whereas the reverse selection does not.

1.1 Related Work

Resilience research holds that redundancy protects a system only when components performing the same function respond differently to perturbation (Elmqvist et al., 2003; Holling, 1973; Yachi & Loreau, 1999). In model populations, organisational labels do not supply that independence. Models from different providers select the same wrong answers above pair-specific chance (Goel et al., 2025; Kim et al., 2025), multi-agent safety work catalogues the population-level failure modes this creates (Franklin et al., 2026; Hammond et al., 2025; Tomašev et al., 2026), and the Hugging Face incident supplies a deployed instance (Greenblatt et al., 2026). The Artificial Hivemind measures cross-model similarity through embeddings and supplies our semantic baseline, but shared meaning is not shared generative organisation (Jiang et al., 2025). AIT offers a feature-free alternative in which NCD approximates the ideal shared-description relation with compressed lengths (Cilibrasi & Vitányi, 2005; Vitányi et al., 2009). Whether that relation predicts correlated failure beyond semantic distance and capability is the question we test. Appendix F extends this discussion.

2 MEASURING GENERATIVE-PROCESS DIVERSITY AND CORRELATED FAILURE

2.1 PROBLEM FORMULATION AND COMPRESSION DISTANCE

Let m_i, $i \in \{1, \ldots, M\}$, be a black-box deployed model configuration and x_{iqr} its r-th response to prompt q. We seek a pairwise statistic $d(i, j)$ that compares deployed configurations through strings generated under matched prompting conditions. Because finite observations do not identify a unique mechanism, "inferred generative-process diversity" denotes an output-derived relation among deployed configurations, not recovered neural computation. It can reflect persistent regularities introduced by model weights, post-training, prompting, decoding, or their interaction.

For compressed length $C(\cdot)$ and concatenation xy, NCD is

$$\mathrm{NCD}_C(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}. \tag{1}$$

NCD is the $p = \infty$ member of a family V_p^C that combines the two directional compression increments, retaining only the larger one. We use V_1^C and V_2^C, which also retain the smaller increment, to test whether results depend on this aggregation choice; Appendix B gives the construction and its theoretical properties. Lower values indicate more shared compressible structure. Reported distances use PPMd variant I through `pyppmd` at library-default memory settings (Miura, 2026). Responses are literal UTF-8 byte strings. No token clustering, language-model symbolisation, or other learned representation enters the compression path. For each prompt and unordered model pair, we average position-paired response distances over $K = \min(R_i, R_j, 50)$ pairs, then average prompts with equal weight. The permutation residualisation below is applied separately to each member of the family.

2.2 RESPONSE CORPUS AND PERMUTATION RESIDUAL

We estimate the compression and semantic predictors from a fixed corpus of repeated responses to Infinity-Chat prompts, generated independently of the benchmark outcomes. Appendix A.1 gives the corpus composition, sampling settings, analysed text field, and filtering procedure.

Raw NCD responds to marginal byte frequencies as well as order. For every response, we generate $P = 20$ deterministic random byte permutations. Each surrogate preserves length and the exact byte multiset while destroying the original order. Let d_{ijq}^{NCD} denote the mean raw NCD for model pair (i, j) on prompt q, and let d_{ijq}^{perm} denote the corresponding mean after applying the same distance and aggregation procedure to the permuted responses. For each prompt, we regress the raw distances on their permutation controls across the $\binom{M}{2}$ unordered model pairs as follows.

$$\begin{aligned} (\widehat{\alpha}_q, \widehat{\beta}_q) &= \arg\min_{a,b} \sum_{1 \le i < j \le M} \left(d_{ijq}^{\mathrm{NCD}} - a - b d_{ijq}^{\mathrm{perm}}\right)^2, \\ \widehat{\varepsilon}_{ijq} &= d_{ijq}^{\mathrm{NCD}} - \left(\widehat{\alpha}_q + \widehat{\beta}_q d_{ijq}^{\mathrm{perm}}\right), \qquad d_{ij}^{\mathrm{IGP}} = \frac{1}{Q} \sum_{q=1}^{Q} \widehat{\varepsilon}_{ijq}. \end{aligned} \tag{2}$$

The projection separates the raw distance into a component explained by the frequency-preserving control and an order-specific remainder, averaged with equal weight across prompts. The residual does not remove every low-order statistic or isolate the mechanism underlying the association. Because d^{IGP} is signed and need not satisfy metric axioms, "distance" below is operational shorthand for pairwise separation.

2.3 SEMANTIC BASELINE

Each response is embedded with `all-MiniLM-L6-v2` and ℓ_2-normalised (Reimers & Gurevych, 2019). Semantic similarity is the mean cross-response cosine similarity for a model pair within a prompt, averaged over prompts; semantic distance is $\mathrm{SemD} = 1 -$ similarity. This follows the embedding-cosine construction used to study the Artificial Hivemind on the same prompt taxonomy (Jiang et al., 2025). The semantic representation is a comparator only and is never used to symbolise

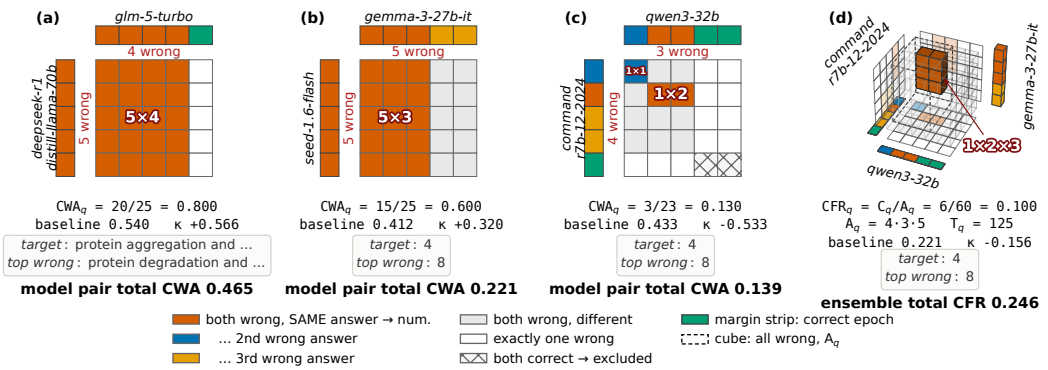

Figure 1: **Correlated wrong-answer agreement and convergent failure rate on GPQA-Diamond.** (a)–(c): each grid crosses five sampled epochs from the row and column models. Coloured cells contribute to the numerator when both models are wrong with the same answer, grey and one-wrong cells enter only the denominator, and hatched both-correct cells are excluded. Item-level raw rate, per-question leave-pair-out baseline, and chance-corrected κ are shown below each grid. Examples span fixed quantiles of pooled pair-level CWA among 700 non-sibling pairs and illustrate why raw agreement must be read against item-specific distractor attraction. (d): Answers in (b)&(c) extended to a three-model ensemble. Each voxel is one triple of epochs, one from each member ($5^3 = 125$ triples). The dashed box is the set of triples in which all three members are wrong ($A_q = 4{\cdot}3{\cdot}5 = 60$), the solid box the subset that agree on one wrong answer ($C_q = 1 \cdot 2 \cdot 3 = 6$), and the statistic is $C_q/A_q = 0.100$ against a leave-ensemble-out hypergeometric baseline of 0.221 ($\kappa = -0.156$). Adding a member can only remove triples from the convergent set. Three member ensemble is shown for legibility, analysis consider 5 member ensembles ($5^5 = 3{,}125$ epoch tuples per item.)

strings for NCD. To complete the comparison with the Artificial Hivemind analysis, Appendix B.6 also characterises variation among repeated responses from the same model under the common prompting and sampling policy.

2.4 Epoch-native correlated failure

The outcome panel comprises ten benchmark families: TruthfulQA, MMLU-Pro, WorldSense, BBEH-mini, GSM8K, AIME, MuSR, GPQA-Diamond, Humanity's Last Exam, and AGIEval (Lin et al., 2022; Cobbe et al., 2021; Rein et al., 2024; Wang et al., 2024; Sprague et al., 2024; Zhong et al., 2024; Center for AI Safety et al., 2026; Kazemi et al., 2025; Benchekroun et al., 2023). Each model answers every available item in five independently sampled epochs under the same generation policy. Multiple-choice letters and numeric answers are normalised by benchmark-specific evaluators; unparseable outputs are tracked separately.

For question q, let $c_{iq}(a)$ count parsed epochs in which model i gives answer a, let $n_{iq} = \sum_a c_{iq}(a)$, and let y_q be the reference answer. Pairwise correlated wrong-answer agreement pools epoch cross-products as follows.

$$\mathrm{CWA}_{ij} = \frac{\sum_q \sum_{a \neq y_q} c_{iq}(a) c_{jq}(a)}{\sum_q \left[n_{iq} n_{jq} - c_{iq}(y_q) c_{jq}(y_q) \right]}. \tag{3}$$

The numerator counts pairings in which both models are wrong with the same answer. The denominator counts every pairing in which at least one model is wrong; it excludes only both-correct pairings. Crossing five epochs per model yields up to 25 pairings per item and retains stochastic answer structure that modal aggregation discards (Figure 1).

Raw agreement depends on the answer space and on how strongly an item's distractors attract the model population. The baseline used in the primary analysis is therefore estimated from the remaining $M - 2$ models separately for each pair under test. Let p_i and p_j denote the pair's respective wrong-answer rates, and let $\bar{h}_{-ij}$ denote the mean question-specific probability that a pair of wrong epochs sampled from the remaining models select the same answer. The expected agreement under

independence and its chance-corrected form are

$$\mathrm{CWA}_{ij}^{\exp} = \frac{p_i p_j}{p_i + p_j - p_i p_j} \bar{h}_{-ij}, \qquad \kappa_{ij} = \frac{\mathrm{CWA}_{ij} - \mathrm{CWA}_{ij}^{\exp}}{1 - \mathrm{CWA}_{ij}^{\exp}}. \tag{4}$$

The first factor in $\mathrm{CWA}_{ij}^{\exp}$ is the probability that both models are wrong conditional on at least one being wrong. This per-question-to-pair procedure is used throughout, with a pair-empirical marginal baseline retained as a sensitivity analysis. Appendix C gives the full construction.

2.5 Association and uncertainty

For each benchmark, the primary statistic is the partial Spearman correlation between d^{IGP} and κ, controlling simultaneously for semantic distance, pair capability level $(a_i + a_j)/2$, and capability gap $|a_i - a_j|$. Rank-transforming before residualisation accommodates monotone capability relationships. The 703 pairs form a complete dyadic network on 38 model nodes. Any two observations that share a model also share model-specific structure in their predictors, capabilities, and failure outcomes, so the effective sampling structure is organised around models rather than pair rows (Aronow et al., 2015). We therefore take the deployed model configuration as the resampling unit. In each of 2,000 node-bootstrap replicates, we sample 38 model slots with replacement, form all unordered pairs of distinct slots, and recompute the partial Spearman correlation on the induced dyads. Repeated model slots reproduce their full incidence pattern across pairs, while self-pairs are omitted. The 2.5th and 97.5th percentiles give the reported interval. Cross-benchmark summaries weight the ten benchmark families equally and report a t interval over the benchmark-level estimates. The capability-plus-semantic partial Spearman correlation with model-node resampling defines the primary analysis; alternative interpolants and chance baselines are used in sensitivity analyses. Ensemble-level cross-benchmark statements are t-intervals over nine source benchmarks; within-benchmark intervals resample selected ensembles and are descriptive only (Appendix D.6).

2.6 Ensemble selection

For an ensemble S of $k = 5$ models, convergent failure is the event that every member is wrong and all give the same wrong answer. With all five epochs per member retained, let A_q count the epoch tuples on item q in which every member is wrong and C_q those that additionally agree (Figure 1). Chance is estimated per item from the wrong epochs of the other $M - k$ models, N_q in total with n_x on wrong answer x, as the hypergeometric probability that k draws coincide. The conditional convergent-failure rate (CFR), its expectation, and the chance-corrected form are

$$\mathrm{CFR}_S = \frac{\sum_q C_q}{\sum_q A_q}, \qquad \mathrm{CFR}_S^{\exp} = \frac{\sum_q A_q\, b_q}{\sum_q A_q}, \quad b_q = \frac{\sum_x \binom{n_x}{k}}{\binom{N_q}{k}}, \qquad \kappa_S = \frac{\mathrm{CFR}_S - \mathrm{CFR}_S^{\exp}}{1 - \mathrm{CFR}_S^{\exp}}. \tag{5}$$

Observed and expected rates share the same items and tuple weights (Appendix D.2). We enumerate all $\binom{38}{5} = 501{,}942$ ensembles per benchmark and keep those with at least 30 estimable items, which admits nine source benchmarks and excludes GSM8K, on which too few ensembles fail five ways.

Because d^{IGP} and semantic distance are collinear ($\rho = 0.789$, Figure 7), ranking ensembles on one measure largely ranks them on the other, so we decouple the two by construction. Each ensemble is scored on both measures by the Vendi score (Friedman & Dieng, 2023), the effective rank of its 5×5 principal submatrix of a fixed positive-semidefinite kernel built from each model's distance profile over the fleet (Appendix D.3). We divide eligible ensembles into five percentile bands of one measure's mean pairwise distance and, within each band, take the top 200 by the other measure's Vendi score. Each selection is compared with random ensembles from the same band, within 0.1 population SD on the held distance, and matched on mean member accuracy and mean pairwise accuracy gap. The reported contrast is the selected-minus-control difference in κ_S, so negative values mean the selected ensembles converge on the same wrong answer less often at fixed capability and fixed level of the other measure. We report the mean over the nine source benchmarks with a 95% t-interval; the difference-form outcome and alternative selectors are in Appendix D.7.

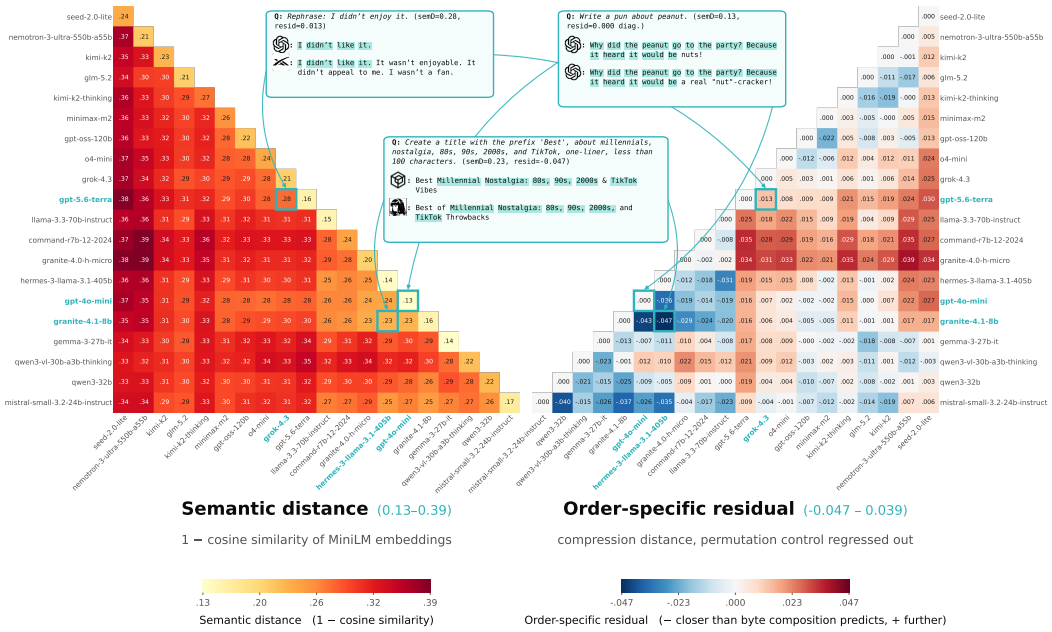

Figure 2: **Semantic distance and the order-specific compression residual for the same responses.** Left: semantic distance. Right: the residual obtained by regressing raw NCD on the within-prompt byte-permutation control and averaging across prompts; positive values are pairs further apart in sequential organisation than byte composition predicts, negative values closer. 20 of 38 models are shown; reported associations use all 703 pairs. Annotated boxes contain response excerpts for a positive, near-zero, and negative residual.

3 Experiments and Results

All compression and semantic predictors are estimated from open-ended responses to Infinity-Chat evaluation (Jiang et al., 2025) and all failure outcomes use disjoint closed-form benchmark responses. This separates the measured texts, although the same models generate both and capability controls derive from benchmark correctness. We observe population structure captured by measures, then test the conditional association with correlated failure, and finally test whether the signal survives in five-model ensembles at fixed semantic distance.

3.1 Semantic and compression-based population structure

Figure 2 places semantic distance beside the order-specific residual for the same responses. Semantic distance records whether outputs mean the same thing and occupies a narrow band of low distance (similar meaning), $[0.159, 0.388]$ across all 703 pairs. The residual records how outputs are sequentially organised, and the two disagree on individual pairs. Exemplar annotations show GPT-4o-mini–Granite-4.1 responses reuse one peanut-pun structure and sit at a near-zero residual. The Granite–Hermes and GPT-5.6–Grok pairs have raw NCD values of 0.675 and 0.678, indistinguishable within the observed range $[0.484, 0.859]$ (Figure 8), yet residuals of -0.047 and $+0.013$: two closely aligned one-line titles are closer than their byte composition predicts, and two responses that share meaning but differ in elaboration are further apart. Residualisation exposes this structure, but it does not simply extract what semantics miss. The residual's rank association with semantic distance is 0.789, higher than raw NCD's 0.662 (Figure 7), so its non-redundant value must be tested against failure with semantic distance held fixed. Subsequent analyses use the residual, with positive values denoting pairs further apart than byte composition predicts.

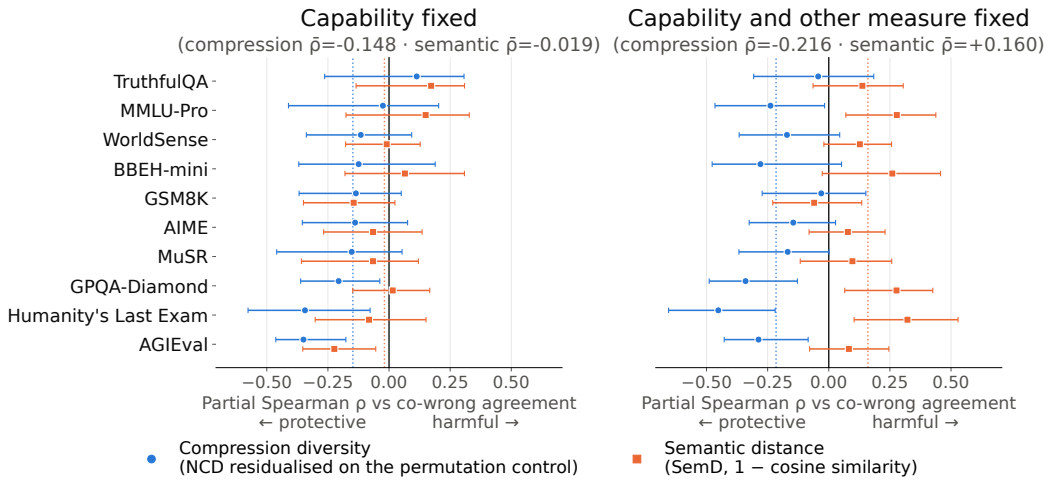

Figure 3: **Dissociation between compression diversity and semantic distance.** Left: each measure with capability level and gap fixed. Right: each measure with capability and its rival fixed. Compression diversity is the order-specific NCD residual. Bars are model-level bootstrap intervals; headings and dotted lines give cross-benchmark means. The relevant test is the paired within-benchmark contrast computed inside each node resample, not overlap of the two marginal intervals. The positive semantic coefficient on the right describes only the residual component of semantic distance surviving a strongly collinear control.

3.2 Conditional associations with correlated failure

Figure 3 compares the order-specific residual with semantic distance under matched control sets. With capability fixed, compression diversity is associated with less correlated failure (-0.148, $[-0.245, -0.050]$), while semantic distance is unresolved (-0.019, $[-0.108, +0.070]$). Holding the rival measure fixed strengthens the compression estimate to -0.216 $[-0.309, -0.122]$; the surviving component of semantic distance is $+0.160$ $[+0.074, +0.246]$.

Semantic distance and the order-specific compression residual are strongly collinear ($\rho = 0.789$, Figure 7), and only about 37% of semantic rank variance remains after controlling for compression. The positive cross-controlled semantic estimate therefore describes this remaining component, not the standalone association of semantic distance with correlated failure. Under cross-control, the compression estimate is negative on all ten benchmarks and its model-node bootstrap interval excludes zero on four, with the largest effects on Humanity's Last Exam (-0.452) and GPQA-Diamond (-0.340). The semantic estimate is positive on nine benchmarks. GSM8K is the exception at -0.060, with its interval spanning zero. The benchmark-paired difference between the two cross-controlled partial correlations has mean -0.375 and 95% t-interval $[-0.545, -0.206]$; Appendix E.3 gives its construction and benchmark-level estimates. Binning pairs into percentile bands of each measure gives the same picture (Appendix C.3). Mean κ is flat across semantic-distance bands, 0.141 to 0.139, whereas across residual bands it falls from 0.167 to 0.086 and is still declining in the top band.

3.3 Compression-selected ensembles fail together less often

Figure 4 reports the difference in CFR κ between selected ensembles and capability-matched random ensembles. Under free selection (Figure 4a), ensembles chosen to maximise compression Vendi score show more negative Δ CFR κ than matched random (-0.047 $[-0.095, +0.000]$), although this arm is dominated by Humanity's Last Exam (Appendix D.7). Ensembles chosen on semantic Vendi score have less negative Δ CFR κ ($+0.018$ $[-0.054, +0.089]$). Because the two measures are collinear, free selection on one also moves the other. Figure 4b shows the Δ CFR κ of each measure when the other measure's mean distance is held inside a percentile band. Realised imbalance on the held distance is below 0.01 SD and on capability below 0.2 SD in every cell (Appendix D.5). With semantic distance held, compression selection lies below matched random in every band (means -0.017 to -0.026),

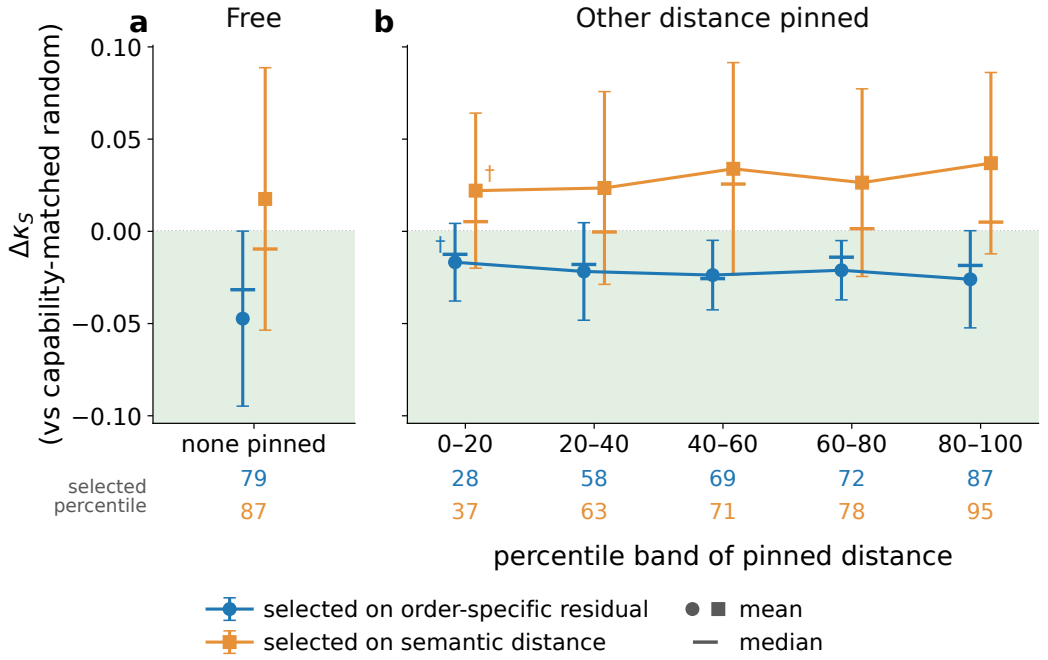

Figure 4: **Vendi selection free and pinned to a percentile band of the other distance.** Each point is the difference in κ_S (Equation 5) between the top 200 five-model ensembles by Vendi score of one measure and capability-matched random ensembles, averaged over nine benchmarks. Compression is the order-specific residual. (a) Selection of ensembles on compression is associated with lower CFR κ than selection on semantic. (b) Selection within the other measure's mean distance pinned to each percentile band shows compression is associated with lower CFR κ across all semantic bands, while semantic has higher mean CFR κ across all compression bands and all intervals include zero. Bars are 95% t-intervals and horizontal marks are medians. Controls match mean accuracy and accuracy gap, and in (b) lie in the same band within 0.1 SD of the pinned distance. Coloured numerals are the median percentile of selected ensembles on selected measure. Daggers mark bands with at least one selection with < 50 controls.

and the interval excludes zero in the 40–60 and 60–80 bands (-0.024 $[-0.043, -0.005]$ and -0.021 $[-0.037, -0.005]$). With compression distance held, semantic selection lies above matched random in every band (means $+0.022$ to $+0.037$), although every interval includes zero and the medians lie closer to zero. Across bands the compression selections span from the 28th percentile of compression distance upward (Figure 4b), so the contrast covers most of the observed range. The interpolants V_1 and V_2 give the same pattern (Appendix D.7).

4 Discussion

We develop inferred generative-process diversity as a black-box measure of model diversity in terms of how models generate observed outputs rather than what those outputs mean. Estimated from single-turn responses to open-ended chat prompts (Jiang et al., 2025), the measure predicts which model pairs choose the same wrong answer on GPQA-Diamond, AIME, and eight other benchmarks beyond semantic similarity and capability (Figure 3). Furthermore, five-model ensembles selected by maximising the Vendi score of the measure at fixed semantic distance converge on one wrong answer less often than capability-matched random ensembles (Figure 4). The ability of generative-process diversity measured on a separate prompt taxonomy to predict correlated failure on ten benchmarks points toward it capturing a persistent property of generation rather than of the specific tasks.

Residualisation of NCD against the byte-permutation control appears critical to isolating failure-relevant signal. Raw NCD is weakly associated with more correlated failure, its frequency-preserving

component (permutation control signal) more strongly, and only the order-specific residual reverses this association showing less correlated failure (Figure 9). This separation parallels the distinction between statistical and algorithmic structure (Zenil et al., 2017), with byte frequencies describing the former and shared sequential organisation the latter. One possible explanation is that byte frequencies reflect length, vocabulary, and formatting conventions that models share for reasons unrelated to how they solve problems, and that many distinct processes can produce the same frequencies.

Semantic distance and the residual are strongly collinear (Figure 7). With capability controlled, semantic distance is not associated with correlated failure, and the component that remains once the residual is also controlled is associated with more of it (Figure 3). This is consistent with semantic distance carrying failure-relevant information only through the variation it shares with the residual. Ensembles selected on the residual at fixed semantic distance converge less than matched ensembles, whereas ensembles selected on semantic distance at fixed residual distance do not (Figure 4). This indicates assessments of model homogeneity based on output meaning (Jiang et al., 2025) do not by themselves capture the variation associated with independent failure. For the tested fleet, selecting model ensembles on semantic diversity was not shown to be harmful under the Vendi selector. Because this rests on the collinearity between the two measures, which need not hold for other fleets, it may not be true in general. As fleets grow, the two measures may come apart, and knowing which one carries the failure-relevant signal becomes important for ensemble design and safety.

The association between inferred generative-process diversity and correlated failure is modest partly because there is little diversity to exploit. If model populations were genuinely diverse one would expect to find both models that produce very different outputs and models that produce similar outputs through different processes. Across 20 providers very little of either is observed. Semantic distance occupies a narrow range (Figure 2), and the process diversity within it is modest. Our intuition is that both reflect convergence in training data, architectures, pipelines and potentially distillation between models. If the resilience framing is taken seriously, and the Hugging Face incident is read as an instance of what a near-monoculture can produce (Greenblatt et al., 2026), this narrowness is itself a risk. If we wish to exploit diversity to improve the performance and resilience of collective systems (Galesic et al., 2023; Tieman et al., 2025; Holling, 1973; Scheffer et al., 2012), a greater number of diverse models are required, along with measures that assess diversity in failure-relevant terms, which the present method provides from outputs alone.

Robust validation of the introduced measure through model-node resampling, cross-controlled specifications, and sensitivity analyses over interpolants, permutation counts, and compressors is performed, however several limitations remain that further analysis can address. Generation cost constrained sampling depth and benchmark breadth, so failure outcomes rest on five epochs per benchmark, and several benchmarks were saturated for the stronger models, which limits usable CWA and CFR and attenuates associations toward zero. CWA combines joint error with agreement on the chosen wrong answer, and separating the two would show whether the residual captures shared blind spots, shared distractor attraction, or both. Natural next steps are to apply the measure to action–observation and tool-use traces, treating model, harness, and tools as one input–output unit, and to use the eigenstructure of the ensemble kernel rather than a single Vendi score to examine how diversity is distributed across an ensemble.

5 Conclusion

Multi-model systems rely on redundancy for robustness, and ecological resilience research indicates that diversity must be judged against the function and failure modes of the system. We introduced inferred generative-process diversity, estimated as the order-specific residual of NCD after a byte-permutation control, as an output-based comparison of the processes that produce model behaviour. Across 38 models it predicts correlated failure on ten benchmarks beyond semantic distance and capability, and ensembles selected on it at fixed semantic distance converge on the same wrong answer less often than capability-matched ensembles. It offers a black-box approach to auditing effective redundancy in model populations, with extension to agent traces and multi-agent systems as the next step.

Reproducibility statement

Section 2 defines every statistic used in the main text, and the appendices give additional detail required to recompute them. Appendix A describes the response corpus, sampling settings, filtering, benchmark panel, question caps, and pair estimability. Appendix B gives details on the compressor, the permutation residualisation, and robustness checks over compressors, permutation counts, and aggregation. Appendix C constructs the pairwise outcome and its chance baselines. Appendix D specifies ensemble enumeration, the ensemble outcome, the kernel and Vendi score, pin bands, capability matching, inference, and full results tables and figures. Appendix E details the partial rank correlations, model-node bootstrap, and paired comparisons. All generation used public model endpoints under a fixed sampling policy, and benchmark evaluation used the Inspect AI framework.

AI use statement

In this work, we used generative AI tools to provide feedback on research methodology and experiments, implement methods, formulate mathematical claims, and interpret results. We have not used generative AI tools to generate synthetic data sets (the analysed model outputs are the object of study), help develop theoretical models or conceptual frameworks, or propose or refine hypotheses, and assistance with translation and support for qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create or edit code, create or modify scientific figures, assist in drafting parts of the paper, edit the paper to improve readability, and suggest a structure for parts of the paper. We have reviewed all AI-assisted work. The authors reviewed and tested all AI-assisted code and checked reported numbers against the analysis outputs. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Peter M. Aronow, Cyrus Samii, and Valentina A. Assenova. Cluster-robust variance estimation for dyadic data. *Political Analysis*, 23(4):564–577, 2015. doi: 10.1093/pan/mpv018.

Youssef Benchekroun, Megi Dervishi, Mark Ibrahim, Jean-Baptiste Gaya, Xavier Martinet, Grégoire Mialon, Thomas Scialom, Emmanuel Dupoux, Dieuwke Hupkes, and Pascal Vincent. Worldsense: A synthetic benchmark for grounded reasoning in large language models. *arXiv preprint arXiv:2311.15930*, 2023. URL https://arxiv.org/abs/2311.15930.

Manuel Cebrián, Manuel Alfonseca, and Alfonso Ortega. Common pitfalls using the normalized compression distance: What to watch out for in a compressor. *Communications in Information and Systems*, 5(4):367–384, 2005. doi: 10.4310/CIS.2005.v5.n4.a1.

Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649:1139–1146, 2026. doi: 10.1038/s41586-025-09962-4. URL https://doi.org/10.1038/s41586-025-09962-4.

Rudi Cilibrasi and Paul M. B. Vitányi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005. doi: 10.1109/TIT.2005.844059.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Rei Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Thomas Elmqvist, Carl Folke, Magnus Nyström, Garry Peterson, Jan Bengtsson, Brian Walker, and Jon Norberg. Response diversity, ecosystem change, and resilience. *Frontiers in Ecology and the Environment*, 1(9):488–494, 2003. ISSN 1540-9309. doi: 10.1890/1540-9295(2003)001[0488:RDECAR]2.0.CO;2.

Paolo Ferragina and Giovanni Manzini. On compressing the textual web. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pp. 391–400. ACM, 2010. doi: 10.1145/1718487.1718536.

Paolo Ferragina, Raffaele Giancarlo, Valentina Greco, Giovanni Manzini, and Gabriel Valiente. Compression-based classification of biological sequences and structures via the universal similarity metric: Experimental assessment. *BMC Bioinformatics*, 8:252, 2007. doi: 10.1186/1471-2105-8-252.

Matija Franklin, Nenad Tomašev, Julian Jacobs, Joel Z. Leibo, and Simon Osindero. AI Agent Traps, March 2026. URL https://papers.ssrn.com/abstract=6372438.

Dan Friedman and Adji Bousso Dieng. The Vendi Score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=g970HbQyk1.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937. doi: 10.1080/01621459.1937.10503522.

Péter Gács. On the symmetry of algorithmic information. *Soviet Mathematics Doklady*, 15:1477–1481, 1974.

Mirta Galesic, Daniel Barkoczi, Andrew M. Berdahl, Dora Biro, Giuseppe Carbone, Ilaria Giannoccaro, Robert L. Goldstone, Cleotilde Gonzalez, Anne Kandler, Albert B. Kao, Rachel Kendal, Michelle Kline, Eun Lee, Giovanni Francesco Massari, Alex Mesoudi, Henrik Olsson, Niccolo Pescetelli, Sabina J. Sloman, Paul E. Smaldino, and Daniel L. Stein. Beyond collective intelligence: Collective adaptation. *Journal of The Royal Society Interface*, 20(200):20220736, March 2023. doi: 10.1098/rsif.2022.0736. URL https://royalsocietypublishing.org/doi/10.1098/rsif.2022.0736.

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. PMLR, 2025. URL https://proceedings.mlr.press/v267/goel25b.html.

Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. METR blog, August 2026. URL https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-Agent Risks from Advanced AI, February 2025. URL http://arxiv.org/abs/2502.14143. arXiv:2502.14143 [cs.MA].

C. S. Holling. Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics*, 4:1–23, 1973. doi: 10.1146/annurev.es.04.110173.000245.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

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*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-2732. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/754d5a526a5ee5a47220664a0eb92751-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. Big-bench extra hard. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 26473–26501, 2025. URL https://aclanthology.org/2025.acl-long.1285/.

Maurice G. Kendall and B. Babington Smith. The problem of m rankings. *The Annals of Mathematical Statistics*, 10(3):275–287, 1939. doi: 10.1214/aoms/1177732186.

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. URL https://proceedings.mlr.press/v267/kim25e.html.

Bjørn Kjos-Hanssen. Interpolating between the Jaccard distance and an analogue of the normalized information distance. *Journal of Logic and Computation*, 32(8):1611–1623, 2022. doi: 10.1093/logcom/exac069. URL https://doi.org/10.1093/logcom/exac069.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021. doi: 10.1073/pnas.2018340118.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012. doi: 10.1561/2200000044.

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. doi: 10.1023/A:1022859003006.

Hugo Larcher, Adrien Carreira, Raphael G., and Christophe Rannou. Anatomy of a frontier lab agent intrusion: A technical timeline of the july 2026 incident. Hugging Face blog, July 2026. URL https://huggingface.co/blog/agent-intrusion-technical-timeline.

Leonid A. Levin. Laws of information conservation (nongrowth) and aspects of the foundation of probability theory. *Problems of Information Transmission*, 10(3):206–210, 1974.

Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M. B. Vitányi. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004. doi: 10.1109/TIT.2004.838101.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.

Hiroshi Miura. miurahr/pyppmd, August 2026. URL https://github.com/miurahr/pyppmd. original-date: 2021-04-13T23:42:33Z.

OpenAI. The Hugging Face incident and the road ahead, August 2026. URL https://openai.com/index/hugging-face-incident-and-the-road-ahead/.

Edward Raff and Charles Nicholas. An Alternative to NCD for Large Sequences, Lempel-Ziv Jaccard Distance. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pp. 1007–1015, New York, NY, USA, August 2017. Association for Computing Machinery. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098111. URL https://dl.acm.org/doi/10.1145/3097983.3098111.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *Proceedings of the First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Samuel R. P.-J. Ross, Owen L. Petchey, Takehiro Sasaki, and David W. Armitage. How to measure response diversity. *Methods in Ecology and Evolution*, 14(5):1150–1167, 2023. ISSN 2041-210X. doi: 10.1111/2041-210X.14087.

Marten Scheffer, Stephen R. Carpenter, Timothy M. Lenton, Jordi Bascompte, William Brock, Vasilis Dakos, Johan van de Koppel, Ingrid A. van de Leemput, Simon A. Levin, Egbert H. van Nes, Mercedes Pascual, and John Vandermeer. Anticipating Critical Transitions. *Science*, 338(6105): 344–348, October 2012. doi: 10.1126/science.1225244. URL https://www.science.org/doi/10.1126/science.1225244.

Dmitry A. Shkarin. PPM: One step to practicality. In *Proceedings of the Data Compression Conference*, pp. 202–211. IEEE Computer Society, 2002. doi: 10.1109/DCC.2002.999958.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jenyYQzue1.

Ross Tieman, Robert Ackland, Katherine Daniell, and Steven J. Lade. Landscape complexity shapes the role of network density and diversity in collective adaptation under disruption. *Research Square preprint*, 2025. doi: 10.21203/rs.3.rs-8115470/v1.

Nenad Tomašev, Matija Franklin, Julian Jacobs, Sébastien Krier, and Simon Osindero. Distributional AGI Safety, May 2026. URL http://arxiv.org/abs/2512.16856. arXiv:2512.16856 [cs.AI].

Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977. doi: 10.1037/0033-295X.84.4.327.

UK AI Security Institute. Inspect AI: Framework for Large Language Model Evaluations, May 2024. URL https://github.com/UKGovernmentBEIS/inspect_ai.

Paul M. B. Vitányi, Frank J. Balbach, Rudi L. Cilibrasi, and Ming Li. Normalized Information Distance. In Frank Emmert-Streib and Matthias Dehmer (eds.), *Information Theory and Statistical Learning*, pp. 45–82. Springer US, Boston, MA, 2009. ISBN 978-0-387-84816-7. doi: 10.1007/978-0-387-84816-7_3. URL https://doi.org/10.1007/978-0-387-84816-7_3.

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. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.

Frans M. J. Willems, Yuri M. Shtarkov, and Tjalling J. Tjalkens. The context-tree weighting method: Basic properties. *IEEE Transactions on Information Theory*, 41(3):653–664, 1995. doi: 10.1109/18.382012.

Danny Wood, Tingting Mu, Andrew M. Webb, Henry W. J. Reeve, Mikel Luján, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of Machine Learning Research*, 24 (359):1–49, 2023. URL https://www.jmlr.org/papers/v24/23-0041.html.

Shigeo Yachi and Michel Loreau. Biodiversity and ecosystem productivity in a fluctuating environment: The insurance hypothesis. *Proceedings of the National Academy of Sciences*, 96(4): 1463–1468, 1999. doi: 10.1073/pnas.96.4.1463.

Hector Zenil, Narsis A. Kiani, and Jesper Tegnér. Low-algorithmic-complexity entropy-deceiving graphs. *Physical Review E*, 96(1):012308, July 2017. doi: 10.1103/PhysRevE.96.012308. URL https://link.aps.org/doi/10.1103/PhysRevE.96.012308.

Hector Zenil, Narsis A. Kiani, and Jesper Tegnér. Algorithmic Information Dynamics: A Computational Approach to Causality with Applications to Living Systems. In *Algorithmic Information Dynamics: A Computational Approach to Causality with Applications to Living Systems*, pp. 310–321. Cambridge University Press, May 2023. doi: 10.1017/9781108596619.021.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2299–2314, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.149. URL https://aclanthology.org/2024.findings-naacl.149/.

Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977. doi: 10.1109/TIT.1977.1055714.

A Data Generation

Across both the predictor corpus and benchmark evaluations, 29 of the 38 deployed model configurations were served through OpenRouter. Four configurations used the OpenAI API directly, four used the Anthropic API directly, and one used the Google API directly.

A.1 Response corpus

The predictor corpus uses the 100 open-ended prompts from the Infinity-Chat prompts introduced for the Artificial Hivemind study (Jiang et al., 2025). For each prompt, we sampled 50 responses from each of 38 deployed model configurations, giving 190,000 generations. Sampling used temperature 1.0, top-p 0.9, no minimum-p, and a 32,768-token output budget. Compression and semantic analyses use only the stored visible assistant response. The prompt, request metadata, and separately recorded reasoning fields do not enter either representation. To prevent language choice from dominating the byte- and embedding-based comparisons, we excluded responses containing more than 20 CJK ideographs. This removed 75 generations (0.04%) and left 189,925 responses; per-prompt computations use the available responses in each model–prompt cell.

A.2 Benchmark panel

Benchmark generation and logging used the UK AI Security Institute's Inspect AI framework (UK AI Security Institute, 2024). Each selected question was evaluated in five independently sampled epochs under the shared generation policy, and the resulting Inspect logs were converted to a common per-model record format while retaining all five responses.

We distinguish task execution, answer-space-specific analysis, and statistical reporting. At execution time, one benchmark family may comprise several Inspect tasks. AGIEval and WorldSense each expand to six tasks, while MuSR and AIME each expand to three. At analysis time, BBEH-mini is divided into multiple-choice, numeric, and other short-answer keys because these answer spaces require different equivalence rules. This gives 12 matrix keys for the ten benchmark families. For reporting, the sufficient counts from the three BBEH-mini keys are pooled before CWA and its baseline are calculated, yielding one BBEH-mini estimate and preventing that benchmark from receiving three votes in cross-benchmark summaries. The reporting panel therefore contains the ten benchmark families used in the main text. BBH is not included because BBEH was designed as its harder successor.

Question caps were applied per Inspect task for API cost reduction. They were 800 for GSM8K, 750 for MMLU-Pro, 300 for HLE-MC, 150 for each WorldSense task, 117 for each MuSR domain, and 100 for each AGIEval task. AIME, BBEH-mini, GPQA-Diamond, and TruthfulQA used their full configured datasets.

A.3 Model population and estimability

The 38-model population gives $\binom{38}{2} = 703$ candidate pairs. Under the requirement of at least 30 questions with a nonzero either-wrong denominator, 703 pairs are estimable on seven keys, 701 on WorldSense, 699 on GPQA-Diamond, and 562 on AIME. AIME is limited by its 90-question cap. Predictor and outcome texts are disjoint, but the same models generate both and capability controls come from the correctness matrices used for the outcome.

B Compression Distance

B.1 Compressor choice and robustness

All reported distances use PPMd variant I through `pyppmd` at library-default memory settings. PPMd is a sequential statistical compressor. It maintains variable-order byte contexts, updates its conditional next-byte distribution as the stream is read, and encodes the resulting predictions with a range coder (Shkarin, 2002). Context Tree Weighting (CTW) is also sequential, recursively weighting bounded-memory tree sources (Willems et al., 1995). By contrast, gzip and LZMA belong to the Lempel–Ziv dictionary family, whose reuse of information across a concatenation is expressed

principally through references to matching phrases (Ziv & Lempel, 1977; Ferragina & Manzini, 2010). Gzip additionally has a finite 32-kB history window, which can make NCD depend on object length once relevant content falls outside that window (Cebrián et al., 2005). PPM compressors have performed consistently among the strongest compressors in compression-based classification, whereas gzip provides a faster but sometimes less discriminative alternative (Ferragina et al., 2007).

For an adaptive compressor, the information transferred across a concatenation is directional. After reading x, the incremental code length of y is $C(xy) - C(x)$. We express its reduction relative to coding y alone as

$$g^C_{x\to y} = 1 - \frac{C(xy) - C(x)}{C(y)} = \frac{C(x) + C(y) - C(xy)}{C(y)}, \qquad T_C(x, y) = \frac{g^C_{x\to y} + g^C_{y\to x}}{2}, \quad (6)$$

where $g^C_{y\to x}$ is defined analogously from $C(yx)$. PPMd and CTW transfer an adaptive conditional context model from the first sequence to the second; the dictionary compressors transfer a phrase dictionary. Both mechanisms can therefore be directional, and $C(xy)$ need not equal $C(yx)$. The symmetric gain T_C is used only for the diagnostic below; the reported NCD matrices retain the fixed orientation in Equation 1.

We first compared PPMd with byte-level CTW (alphabet size 256 and depth 12), LZMA, and gzip on raw model responses. Because CTW is costly on byte strings, this diagnostic uses a deliberately heavy subsample comprising 16 of the 38 models, 15 of the 100 prompts, and one response position per prompt shared across models. Responses are not truncated; four model–prompt responses above the 8-KiB CTW limit were excluded from every compressor. This leaves 1,746 pair–prompt observations and 120 model-pair means. Table 1 reports agreement with the PPMd ordering and finite-length saturation.

Table 1: **Raw-response compressor robustness.** Rank agreement is Spearman's ρ between the 120 model-pair mean distances under PPMd and each alternative compressor. The remaining columns use all 1,746 pair–prompt observations. Absolute NCD levels are compressor-specific and are not a common calibrated scale.

Compressor	ρ with PPMd	Mean NCD	NCD ≥ 0.95
PPMd	—	0.787	2.8%
CTW	0.884	0.932	38.1%
LZMA	0.918	0.672	0.0%
gzip	0.958	0.769	0.7%

The model-pair ordering is therefore broadly preserved under all three alternatives, most closely under gzip. The mean distance itself is not a quality score. LZMA's lower mean, for example, reflects a different finite-length scale. Saturation is the relevant failure mode for discrimination. CTW places 38.1% of observations at or above 0.95, compared with 2.8% for PPMd, leaving much less variation among raw response pairs.

We next separated transfer of sequential statistics from reuse of literal substrings in a paired 2×2 factorial experiment. Each sequence pair either shared or did not share a first-order byte-transition law (*shared context*), and independently contained the same or disjoint randomly generated byte blocks (*shared exact phrases*). The two sequences were generated independently, and shared blocks were inserted at different positions and in different orders. Phrase lengths were 9 bytes at the shortest sequence length, 30 bytes at 245 bytes, and 32 bytes thereafter; phrase coverage was approximately 25%, except at the shortest length where the single block covered 12%. The six sequence lengths—73, 245, 664, 1,302, 2,579, and 4,000 bytes—are the 10th, 25th, 50th, 75th, 90th, and 95th percentiles of all 190,000 responses in the predictor corpus. We used 12 paired replicates at each length. A factorial main effect is the change in T_C when one factor is shared, averaged over the two levels of the other factor.

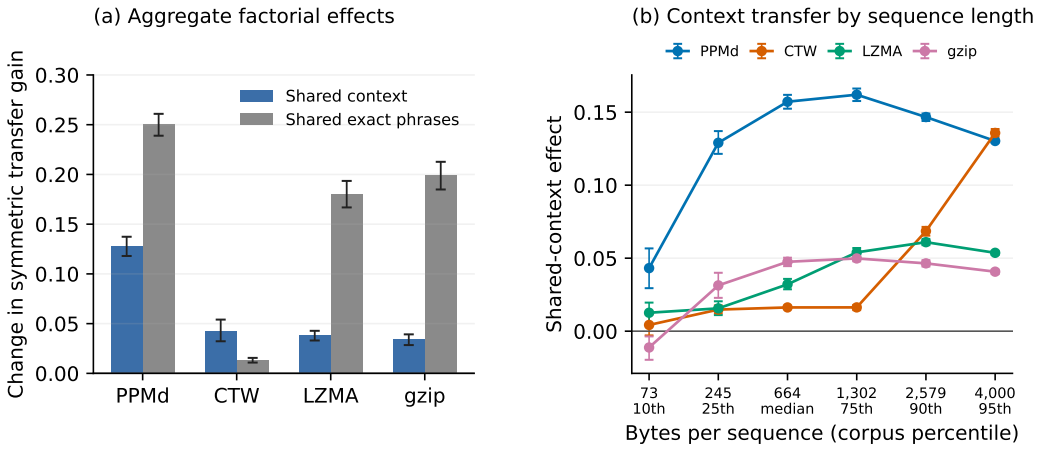

Figure 5: **Sequential-context and exact-phrase transfer.** (a) Main effects of sharing a first-order transition law or exact byte blocks on the symmetric transfer gain in Equation 6. Bars are means over 72 paired units; intervals are 95% bootstrap intervals. (b) The shared-context effect at six sequence lengths drawn from the empirical response-length distribution. Each of x and y has the displayed length, so the concatenation has approximately twice as many bytes. Intervals resample the 12 replicates at each length. PPMd transfers substantially more shared context over the central range of the corpus; CTW approaches it only near the 95th percentile.

PPMd's aggregate shared-context effect is 0.128 (95% interval $[0.118, 0.137]$), compared with 0.043 for CTW, 0.038 for LZMA, and 0.034 for gzip; all three paired PPMd contrasts remain significant after Holm correction ($p = 1.5 \times 10^{-4}$; Appendix E.4). PPMd also has the largest exact-phrase effect (0.250), so the result is not that PPMd ignores literal reuse. Rather, it combines phrase reuse with substantially stronger transfer of non-verbatim sequential statistics than the dictionary compressors. CTW exhibits the strongest context preference relative to its own phrase effect, but both effects are small at typical response lengths. At 1,302 bytes, CTW's context effect is 0.016 against PPMd's 0.162; at 2,579 bytes it is 0.069 against 0.147; only at 4,000 bytes does CTW reach 0.136 against PPMd's 0.130. CTW can therefore detect the controlled transition structure, but requires substantially longer strings to do so under the byte-level configuration used here.

The compressor choice is also coupled to the aggregation protocol. We compress position-paired responses separately and average their distances, rather than concatenate all 50 responses in a model–prompt cell into one stream. The latter construction drove raw PPMd distances towards the upper boundary as cell length increased in our development diagnostics; response-level pairing retained substantially more variation and confines directional transfer to one response pair. Taken together, the preserved model-pair ordering, limited saturation, and finite-length context transfer support PPMd for the present response-level analysis. The comparison is nevertheless descriptive and heavily subsampled. It does not repeat the held-out correlated-failure models under every compressor, which remains future work.

B.2 The V_p family on sets and sequences

The finite-set construction makes the relation among V_1, V_2, and V_∞ exact. For finite sets A and B, let

$$\Delta_p(A,B) = \left(|B \setminus A|^p + |A \setminus B|^p\right)^{1/p}, \qquad V_p(A,B) = \frac{\Delta_p(A,B)}{|A \cap B| + \Delta_p(A,B)}, \tag{7}$$

for $1 \le p < \infty$, with $\Delta_\infty(A,B) = \max\{|B \setminus A|, |A \setminus B|\}$ and $V_p(\varnothing,\varnothing) = 0$. The endpoints are

$$V_1(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}, \qquad V_\infty(A,B) = \frac{\max\{|B \setminus A|, |A \setminus B|\}}{\max\{|A|, |B|\}}. \tag{8}$$

Thus V_1 is the Jaccard distance, V_∞ is the set analogue of Normalised Information Distance (NID), and V_2 uses the Euclidean norm of the two directional differences. In the symmetric Tversky family, these are the endpoints $V_1 = D_{1/2,2}$ and $V_\infty = D_{0,1}$ (Tversky, 1977; Kjos-Hanssen, 2022).

Kjos-Hanssen proves that V_p is a metric for every $p \in [1, \infty]$ (Kjos-Hanssen, 2022). The main step can be seen directly. Set containment gives

$$|B \setminus A| \leq |B \setminus C| + |C \setminus A|, \qquad |A \setminus B| \leq |A \setminus C| + |C \setminus B|.$$

Applying Minkowski's inequality to these two coordinates yields

$$\Delta_p(A, B) \leq \Delta_p(A, C) + \Delta_p(C, B).$$

The normalisation in Equation 7 also preserves the triangle inequality because $|B \setminus A| \leq \Delta_p(A, B)$ supplies the condition required by the ratio-normalisation lemma of Kjos-Hanssen (2022). Non-negativity, symmetry, and identity follow from the two set differences. This proves the metric result without selecting a special value of p.

The ordering follows from the standard ordering of norms on $\mathbb{R}^2$.

$$\Delta_\infty \leq \Delta_2 \leq \Delta_1 \quad \Longrightarrow \quad V_\infty \leq V_2 \leq V_1. \tag{9}$$

The implication holds because $t/(|A \cap B| + t)$ is increasing in t. If one directional difference is zero, all three values coincide. If the differences are balanced at (t, t), their unnormalised values are t, $\sqrt{2}t$, and $2t$ for $p = \infty, 2, 1$, respectively. The members therefore differ most when each object contains substantial information absent from the other.

The set result also suggests a direct route to sequence distances. One can map a sequence s to its Lempel–Ziv phrase dictionary $F(s) = \mathrm{LZSet}(s)$ and pull the set metric back as $V_p(F(s), F(t))$; at $p = 1$, this recovers the Lempel–Ziv Jaccard distance (Raff & Nicholas, 2017; Kjos-Hanssen, 2022). The construction is exact but inherits the representation chosen by F. The map is parser-specific and non-injective, so distinct sequences with the same phrase set receive distance zero. This limitation motivates an object-level interpolation that does not first reduce each sequence to an explicit feature set.

B.3 Extension to strings and finite objects

For binary strings x and y, let $K(x)$ be prefix Kolmogorov complexity, $K(x, y)$ the complexity of a fixed effective pairing, and

$$I_K(x; y) = K(x) + K(y) - K(x, y).$$

Using the directional conditional complexities $k_{x|y} = K(x \mid y)$ and $k_{y|x} = K(y \mid x)$, define

$$\Delta_p^K(x, y) = \left(k_{x|y}^p + k_{y|x}^p\right)^{1/p}, \qquad V_p^K(x, y) = \frac{\Delta_p^K(x, y)}{I_K(x; y) + \Delta_p^K(x, y)}. \tag{10}$$

Let $N = K(x, y, z) + 2$ and $\lambda_N = O(\log N)$. Symmetry of information gives

$$K(x, y) = K(x) + K(y \mid x) + O(\lambda_N) = K(y) + K(x \mid y) + O(\lambda_N)$$

(Levin, 1974; Gács, 1974). Substitution into Equation 10 gives the two endpoint identities

$$V_1^K(x, y) = 1 - \frac{I_K(x; y)}{K(x, y)} + O\left(\frac{\lambda_N}{K(x, y)}\right), \tag{11}$$

$$V_\infty^K(x, y) = \frac{K(x, y) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}} + O\left(\frac{\lambda_N}{\max\{K(x), K(y)\}}\right). \tag{12}$$

We refer to Equation 11 as the *algorithmic Jaccard distance*; Equation 12 is NID (Li et al., 2004). At the conditional-complexity level, $V_1^K = D_{1/2,2}^K$ and $V_\infty^K = D_{0,1}^K$ are exact. Only their reduction to the displayed joint-complexity forms incurs logarithmic slack.

Universality and the information retained at finite p. Let

$$M_K = \max\{K(x \mid y), K(y \mid x)\}, \qquad m_K = \min\{K(x \mid y), K(y \mid x)\}.$$

Then $\Delta_\infty^K = M_K$ and, for finite p, $\Delta_p^K = (M_K^p + m_K^p)^{1/p}$. Norm equivalence in two dimensions gives

$$\Delta_\infty^K \leq \Delta_p^K \leq 2^{1/p} \Delta_\infty^K. \tag{13}$$

After absorbing the $O(\lambda_N)$ possible negativity of I_K into the symmetry-of-information slack, the same comparison holds after normalisation.

$$V_\infty^K(x,y) \le V_p^K(x,y) \le 2^{1/p} V_\infty^K(x,y) + O\left(\frac{\lambda_N}{H}\right), \tag{14}$$

where H is the denominator scale and $2^{1/\infty} = 1$. NID minorises every admissible upper-semicomputable normalised distance d satisfying the density condition (Li et al., 2004; Vitányi et al., 2009). Equation 14 therefore implies

$$V_p^K(x,y) \le 2^{1/p} d(x,y) + O\left(\frac{\lambda_N}{H}\right). \tag{15}$$

Thus V_2^K and V_1^K inherit NID's minorisation property within factors $\sqrt{2}$ and 2, respectively, while only V_∞^K retains the sharp coefficient one. This is a difference in worst-case universality, not an ordering of usefulness for a particular outcome.

Finite p remains sensitive to a quantity absent from NID. Up to logarithmic slack,

$$V_\infty^K = \frac{M_K}{I_K + M_K}, \qquad V_1^K = \frac{M_K + m_K}{I_K + M_K + m_K},$$

and hence

$$V_1^K - V_\infty^K = \frac{I_K m_K}{(I_K + M_K)(I_K + M_K + m_K)} + O\left(\frac{\lambda_N}{H}\right). \tag{16}$$

At fixed shared information $I_K > 0$ and larger directional cost M_K, every finite-p member varies with the smaller cost m_K, whereas NID does not. The members agree when one directional cost vanishes and separate as the two costs become more balanced. Retaining this second direction may be useful when the target depends on reciprocal novelty—neither object can be described from the other by a short program—rather than on the larger one-sided difference alone. This does not imply that a finite-p member must predict correlated failure better. That question is outcome-specific and empirical. Moreover, the universality statements above concern the ideal K-level quantities; a real-compressor approximation, and especially its signed residual after projection, does not inherit them automatically.

The metric proof also transfers, with the same qualification. Conditional descriptions compose according to

$$K(x \mid y) \le K(x \mid z) + K(z \mid y) + O(\lambda_N),$$

and the analogous inequality holds in the opposite direction. Minkowski's inequality therefore gives

$$\Delta_p^K(x,y) \le \Delta_p^K(x,z) + \Delta_p^K(z,y) + O(\lambda_N). \tag{17}$$

For the ratio normalisation, symmetry of information and the same composition inequality imply

$$I_K(x;z) - I_K(x;y) \le K(z \mid y) + O(\lambda_N) \le \Delta_p^K(y,z) + O(\lambda_N).$$

Applying the set-level ratio argument then yields the triangle inequality for V_p^K up to relative $O(\lambda_N/H)$ slack, where H is the scale of the denominators. Identity is approximate rather than literal. Even $K(x \mid x)$ is $O(1)$, and distinct strings related by a fixed reversible procedure can have $K(x \mid y), K(y \mid x) = O(1)$. More precisely, for a family of pairs (x_n, y_n), their mutual description lengths are *uniformly bounded* if there is one constant c, independent of n, such that

$$\max\{K(x_n \mid y_n), K(y_n \mid x_n)\} \le c \quad \text{for every } n.$$

If the denominator scale H_n grows, then $V_p^K(x_n, y_n) = O(1/H_n)$ and tends to zero. Thus V_p^K separates growing algorithmic objects only up to uniformly bounded mutual description length; it is a logarithmic pseudometric rather than an exact metric on literal finite strings.

The same result applies to any finite object supplied with an effective self-delimiting encoding. A computable reversible change of encoding alters Kolmogorov complexity by at most an additive constant, which is absorbed by the logarithmic term. The object-level statement is therefore not tied to one byte representation, although any computable approximation using a real compressor remains representation-dependent.

B.4 COMPRESSED-LENGTH INTERPOLATION

Let $c_x = C(x)$, $c_y = C(y)$, and $c_{xy} = C(xy)$, and define the estimated directional complexities

$$u_C = c_{xy} - c_y, \qquad v_C = c_{xy} - c_x, \qquad I_C = c_x + c_y - c_{xy}.$$

Replacing K by a normal compressor C in Equation 10 gives $\Delta_p^C = \|(u_C, v_C)\|_p$ and $V_p^C = \Delta_p^C/(I_C + \Delta_p^C)$. The normal-compressor axioms—idempotence, monotonicity, symmetry, and distributivity up to $\varepsilon(n) = O(\log n)$—supply the compressor analogue of conditional composition,

$$C(x \mid z) \le C(x \mid y) + C(y \mid z) + \varepsilon(n),$$

so the proof above carries through with $O(\varepsilon(n)/H)$ relative slack (Cilibrasi & Vitányi, 2005). When $u_C, v_C \ge 0$, the endpoints simplify to

$$V_1^C = \frac{2c_{xy} - c_x - c_y}{c_{xy}}, \qquad V_\infty^C = \frac{c_{xy} - \min\{c_x, c_y\}}{\max\{c_x, c_y\}}, \tag{18}$$

so V_1^C is the computable algorithmic Jaccard distance and V_∞^C is NCD. The middle member uses $\Delta_2^C = (u_C^2 + v_C^2)^{1/2}$.

A real compressor need not satisfy the normality axioms exactly. PPMd is left-to-right and $C(xy)$ need not equal $C(yx)$; the reported statistic uses the fixed orientation shown in Equation 1. We therefore treat all three members as empirical compression dissimilarities, not exact metrics. The implementation clips negative directional estimates to zero for V_2 and clips all reported values to $[0, 1]$. Under compressor normality these adjustments are within the same $O(\varepsilon(n)/H)$ approximation.

B.5 RESIDUALISATION AS A WITHIN-PROMPT PROJECTION

For a prompt q, collect the $n = \binom{M}{2} = 703$ raw pair distances into the vector $y_q \in \mathbb{R}^n$ and the corresponding permutation-control distances into $p_q \in \mathbb{R}^n$. Let $\iota_n = (1, \dots, 1)^\top$ and

$$X_q = [\iota_n \quad p_q], \qquad P_q = X_q(X_q^\top X_q)^{-1} X_q^\top. \tag{19}$$

The fitted component and residual are

$$\widehat{y}_q = P_q y_q, \qquad r_q = (I_n - P_q) y_q, \qquad d^{\mathrm{IGP}} = \frac{1}{Q} \sum_{q=1}^{Q} r_q. \tag{20}$$

The matrix P_q is symmetric and idempotent, $P_q^\top = P_q$ and $P_q^2 = P_q$, and therefore projects onto the span of the intercept and the prompt-specific control. The normal equations give

$$X_q^\top r_q = 0. \tag{21}$$

The residuals consequently have mean zero and zero sample covariance with the permutation control within each prompt. Positive residuals are pairs whose observed distance is greater than predicted from their frequency-preserving controls for that prompt; negative residuals are closer than predicted. Zero is a fitted reference point, not the absence of diversity.

The regression is fitted separately for each prompt because pair comparisons are matched on prompt content and the scale of the control varies across prompts. Current slopes range from 1.03 to 7.65, with mean 3.13. Averaging r_q then gives every prompt equal weight. Orthogonality is a within-prompt property. It does not require the averaged residual vector to be exactly orthogonal to the averaged control vector, because cross-prompt products remain. Their pair-level Pearson correlation is -0.048 in the current matrices.

Figure 7 shows what the projection changes at the pair level. The point colour is fixed to raw NCD and the point size to semantic distance in all three panels. Raw NCD is strongly ordered by its own colour, as it must be. The permutation panel retains much of that ordering. After projection, the colours are rearranged vertically. At a fixed semantic distance, pairs with larger raw NCD need not have larger order-specific residuals. The partial Spearman association between raw NCD and the residual, controlling for semantic distance, is -0.409. This conditional reordering is distinct from the positive marginal association between the residual and semantic distance ($\rho_s = 0.789$, against

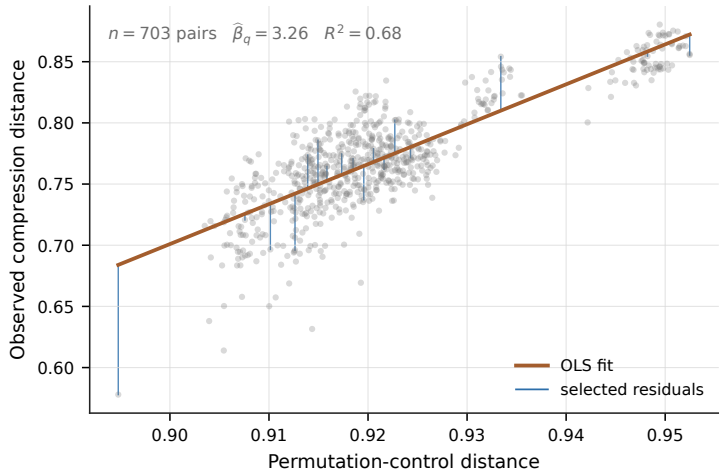

Figure 6: **Within-prompt residualisation.** One representative current prompt with 703 model pairs. The line is the within-prompt OLS fit and the blue segments show a subset of retained residuals.

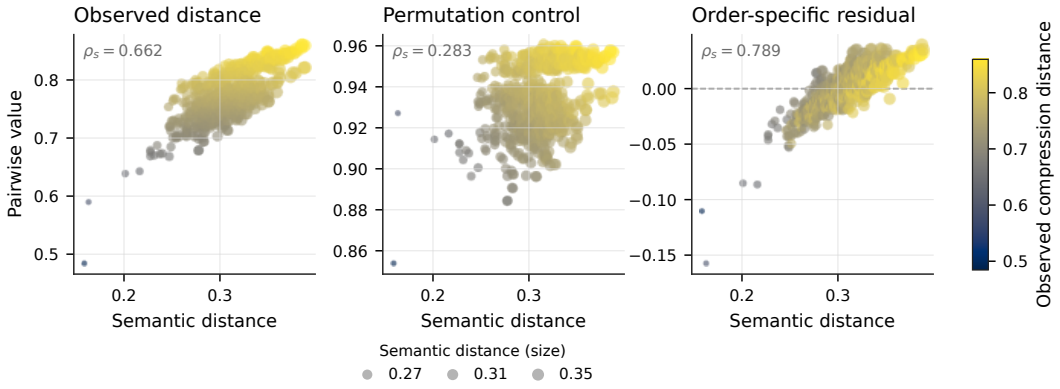

Figure 7: **Pairwise effect of residualisation.** All 703 model pairs are plotted against semantic distance. Point colour records raw NCD and point size also records semantic distance; both encodings are held fixed across panels. The panels show raw NCD, its permutation control, and the retained order-specific residual. Rank associations with semantic distance are printed within each panel.

0.662 for raw NCD, printed in the panels). The residual is therefore strongly collinear with semantic distance: $1 - \rho_s^2 \approx 0.38$, so only about 37% of semantic rank variance is not shared with the residual, which is the component the cross-controlled semantic estimate in Section 3 describes.

Figure 8 shows the raw NCD matrix beside the semantic and residual matrices of Figure 2, with the same annotated pairs. Across all 703 pairs raw NCD spans $[0.484, 0.859]$, against $[0.159, 0.388]$ for semantic distance, and the two agree on coarse population structure but not on many individual pairs (Spearman $\rho = 0.662$, Pearson $r = 0.745$). The Granite–Hermes and GPT-5.6–Grok pairs have raw NCD values of 0.675 and 0.678, indistinguishable within that range, yet residuals of -0.047 and $+0.013$.

Figure 9 decomposes compression distance while holding capability level and gap fixed. The components differ in sign. Raw NCD is mildly associated with *more* correlated failure ($\bar{\rho} = +0.101$), and the frequency-only permutation component is more positive ($+0.150$). Only the order-specific residual is associated with *less* correlated failure (-0.148). Raw compression distance therefore does not merely add noise. It suggests the opposite conclusion. We therefore define compression-derived diversity using the permutation residual. Only after removing the frequency-preserving component does the measure identify pairs with less correlated failure.

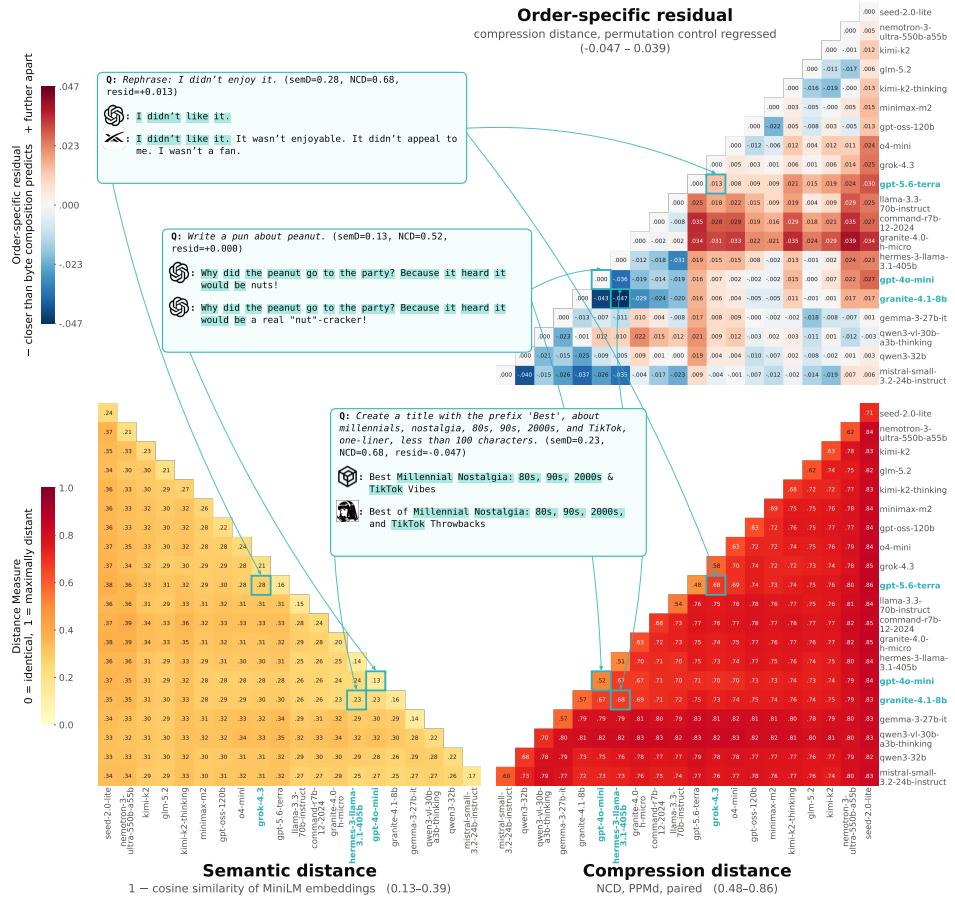

Figure 8: **Semantic, raw-compression, and order-specific views of the same responses.** The lower matrices show semantic distance (left) and raw paired PPMd NCD (right). The upper-right matrix shows the order-specific compression distance obtained by residualising NCD against the within-prompt byte-permutation control and averaging across prompts. 20 of 38 models are displayed; reported matrix associations use the 703 off-diagonal pairs. Annotated boxes contain response excerpts and give the semantic distance, raw NCD, and residual for a positive, near-zero, and negative residual.

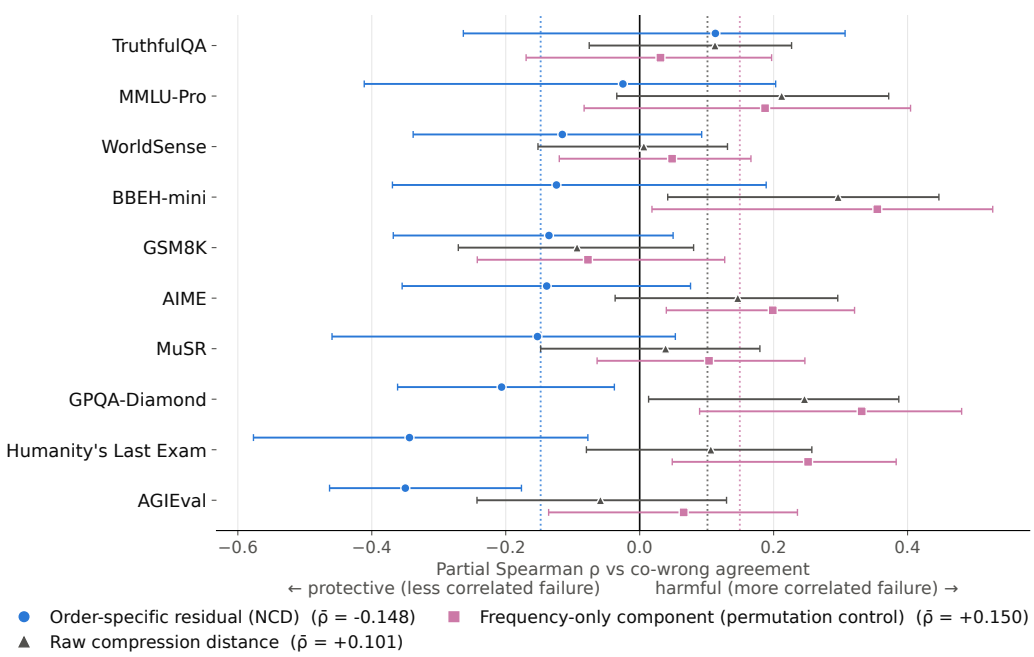

Figure 9: **Compression components against correlated failure.** Points are benchmark-specific partial Spearman correlations with capability level and gap held fixed; bars are model-level bootstrap intervals and dotted lines are cross-benchmark means. Negative values denote less chance-corrected co-wrong agreement. Raw NCD ($\bar{\rho} = +0.101$) and the frequency-only permutation component ($+0.150$) are associated with more correlated failure on average, whereas the order-specific residual is associated with less (-0.148).

B.6 Within-model response variability

The Artificial Hivemind characterises open-ended homogeneity at two levels. These are intra-model repetition among responses repeatedly sampled from one model and inter-model similarity across models (Jiang et al., 2025). Our primary analysis extends the inter-model comparison. To complete the comparison on the same prompt taxonomy, we apply the corresponding within-model semantic construction to our 38-model response corpus and place it beside within-model compression variability. The construction is analogous rather than a numerical replication because we use MiniLM embeddings, but both analyses average pairwise embedding similarity among repeated responses to the same prompt.

This extension uses NCD at a second scale. Between models, NCD compares the sequential organisation expressed by different deployed configurations. Within a model and prompt, it compares repeated draws from the same conditional response distribution and therefore characterises realised stochastic response variability under the fixed decoding policy. It is not an estimator of an intrinsic entropy or randomness parameter. The measured variation includes the effects of the prompt, model, interface, and sampling procedure. Nor does it imply that a model changes its generative process across prompts. Each prompt conditions the same deployed configuration on a different input.

For model m and prompt q, within-model semantic distance d_{mq}^{sem} is one minus the mean cosine similarity over all distinct pairs among the 50 response embeddings. Within-model compression distance d_{mq}^{NCD} pairs response r with response $r + 25$ and averages the resulting 25 disjoint NCD values. The permutation control d_{mq}^{perm} is computed from the same 25 response pairs. Because this is a separate estimand from the cross-model analysis, we fit new coefficients across the 38 within-model

observations for each prompt as follows.

$$(\widehat{\alpha}_q^{\text{intra}}, \widehat{\beta}_q^{\text{intra}}) = \arg\min_{a,b} \sum_{m=1}^{M} \left(d_{mq}^{\text{NCD}} - a - b d_{mq}^{\text{perm}}\right)^2,$$
$$\widehat{\varepsilon}_{mq}^{\text{intra}} = d_{mq}^{\text{NCD}} - \left(\widehat{\alpha}_q^{\text{intra}} + \widehat{\beta}_q^{\text{intra}} d_{mq}^{\text{perm}}\right). \tag{22}$$

The resulting residual is relative to the model population for that prompt. Positive values denote greater within-model order-specific variation than predicted from byte composition, and negative values denote less. As in the pairwise construction, the residual is orthogonal to the permutation control within each prompt; it is not residualised against semantic distance.

Models differ under the shared policy. Mean semantic distance across all model–prompt cells is 0.197, with model means from 0.134 to 0.273; mean NCD is 0.614, with model means from 0.483 to 0.719. Treating prompt as a repeated-measures block, model effects are present for semantic distance (Kendall's $W = 0.357$), raw NCD ($W = 0.538$), and the intra-model order-specific residual ($W = 0.422$); all three Friedman tests have $p < 10^{-250}$. Appendix E.4 defines the coefficient and test. These are differences in the degree of response variability, not evidence of broad semantic disagreement. Repeated responses remain close in the embedding space for every model.

The measures share substantial variation. Across the 38 model means, semantic distance and raw NCD have Pearson $r = 0.848$, Spearman $\rho_s = 0.875$, and $R^2 = 0.719$ under the linear fit in Figure 10. The order-specific residual follows the same broad pattern. Its association with semantic distance across model means is $r = 0.931$ and $\rho_s = 0.917$. This does not indicate a failure of the projection. The permutation control is only weakly associated with semantic distance across model means ($r = 0.267$), so removing the composition-predicted component leaves most of the semantic-associated sequential variation. At the prompt level, the mean Pearson association across models decreases from 0.732 for raw NCD to 0.702 for the residual; averaging over prompts exposes stable model-level differences that both measures track.

The association is strong but not exact. The linear semantic trend leaves 28.1% of between-model NCD variance unexplained, and rankings by the two model means disagree for 115 of the 703 pairwise model orderings. To show where the measures differ, Figure 10 defines $\delta_m = \bar{d}_m^{\text{NCD}} - (\widehat{a} + \widehat{b}\bar{d}_m^{\text{sem}})$, where $(\widehat{a}, \widehat{b})$ is fitted across the 38 model means. DeepSeek-R1-Distill-Llama-70B, Qwen3-VL-Thinking, and Seed-1.6-Flash have more within-model NCD than their semantic distance predicts. Both GPT-5.6-Terra configurations and Mistral-Nemo have less. These departures are descriptive differences between the measures; they do not by themselves establish which component predicts a separate outcome.

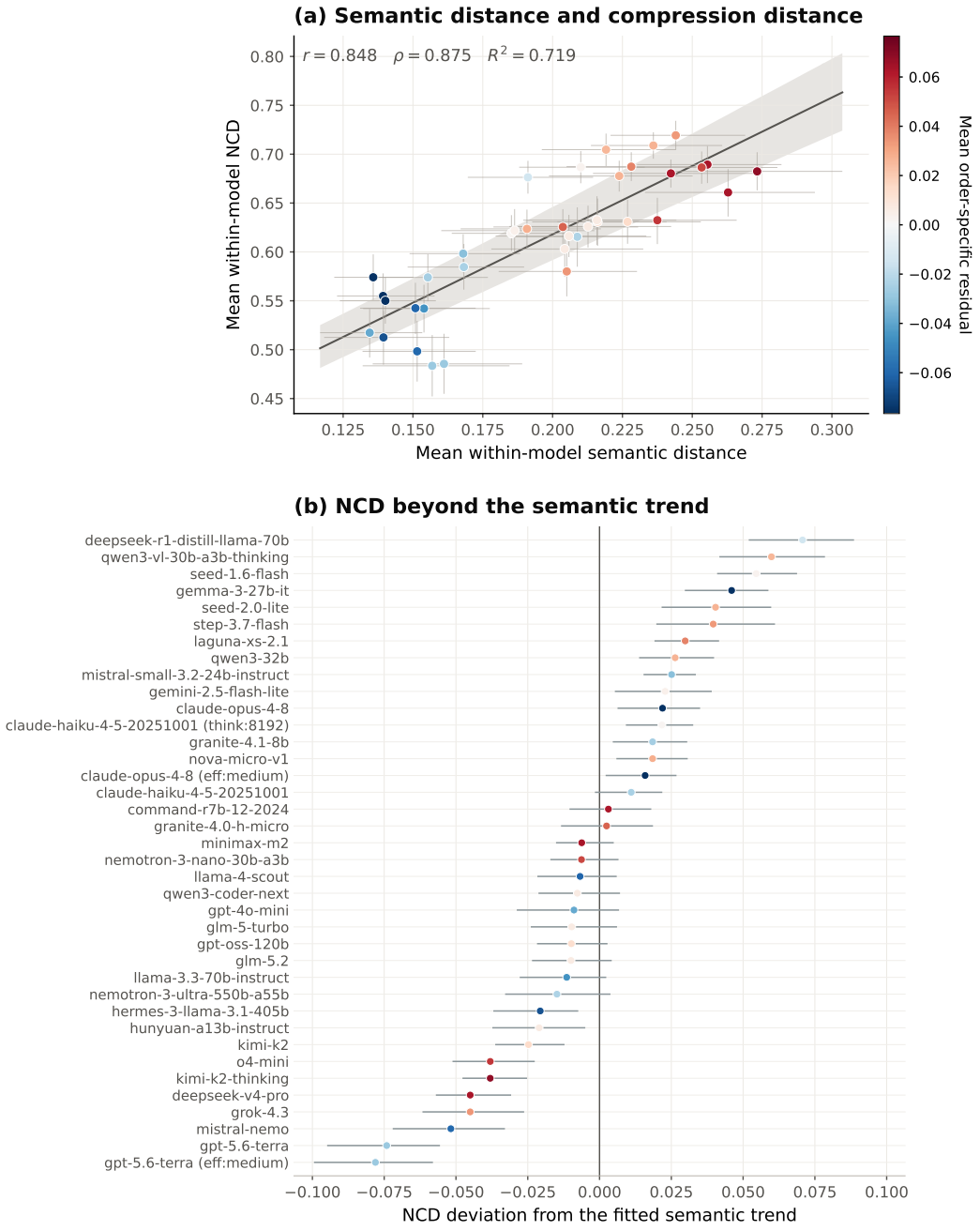

Figure 10: **Within-model semantic and compression variability.** (a) Each point is one model, positioned by its mean semantic distance and mean raw NCD over 100 prompts. Horizontal and vertical bars are 95% prompt-bootstrap intervals. The line is the OLS fit across the 38 model means, the band refits that relationship in each prompt-bootstrap sample, and colour records the separately fitted mean intra-model order-specific residual from Equation 22. (b) Model-specific NCD deviation δ_m from the fitted semantic trend. Intervals resample prompts jointly across models and refit the trend in every replicate. Positive values indicate more compression variability than semantic distance predicts; negative values indicate less. The figure therefore displays both the shared response-variability component and the model-specific departures from it.

The intra-model analysis therefore completes the comparison with the Artificial Hivemind while giving NCD a complementary interpretation. The same output-derived statistic characterises separation between deployed model configurations and stochastic response variability within a configuration. At the within-model scale, semantic and compression variability are related but not interchangeable. Semantic distance records changes in expressed meaning, whereas NCD also responds to how repeated draws vary in their sequential organisation. Accordingly, this analysis characterises how the two measures covary under repeated sampling, but does not independently establish that compression captures information beyond semantic distance. Evidence that compression carries outcome-relevant information beyond semantic distance comes instead from the held-out correlated-failure analysis, where semantic distance and capability are controlled directly. We hypothesise that NCD may capture non-trivial variation in stochastic response variability beyond semantic distance; testing this possibility requires a dedicated analysis and is left to future work.

B.7 Permutation count and aggregation

Each response byte stream is permuted $P = 20$ times with deterministic seeds derived from the master seed, query identifier, and label. A permutation-count sensitivity analysis over $N \in \{5, \ldots, 320\}$ on 400 outputs found 2.5 bits of per-output drift between $N = 20$ and $N = 320$, 0.4% of a typical 655.6-bit residual, with mean drift $+0.15$ bits. For a cell containing 50 outputs, the $N = 20$ control shifted the residual by 0.7 bits and added 0.01 bits of standard error. All tested permutation counts satisfied both adequacy criteria, so $N = 20$ is conservative.

Within a model-pair/query cell, $K = \min(R_i, R_j, 50)$ position-paired responses are averaged. Queries then receive equal weight, so a query contributing 25 response pairs counts as much as one contributing 50. Reconstruction from the per-query cache reproduces the stored matrices exactly.

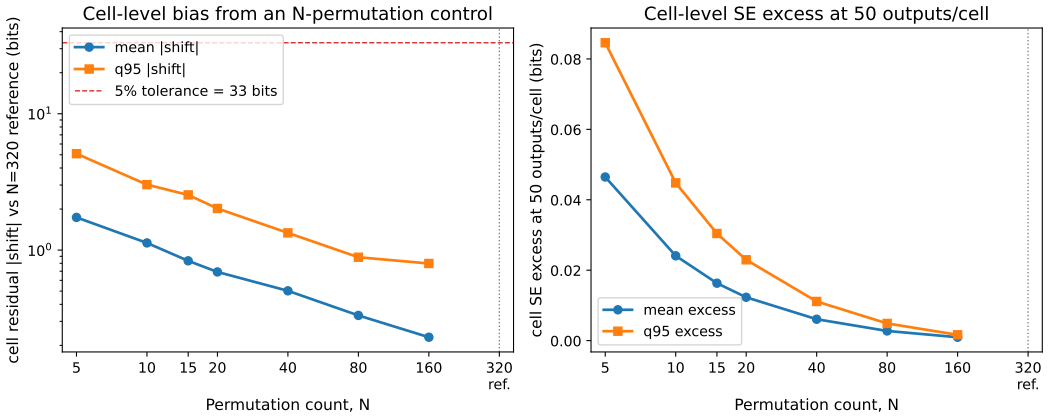

Figure 11: **Permutation-count sensitivity for 50-output cells.** Left: mean and 95th-percentile absolute shifts in cell residuals at each tested permutation count relative to $N = 320$; the dashed line is the 5% tolerance of 33 bits. Right: mean and 95th-percentile excess standard error for a 50-output cell relative to $N = 320$. The $N = 320$ reference is marked on both axes but omitted as a data point because both quantities are zero by construction. At the chosen value $N = 20$, the mean shift is 0.7 bits and the excess standard error is 0.01 bits.

B.8 Robustness within the raw-byte family

The finite-p members test whether retaining the smaller directional compression increment changes the association with correlated failure. This sensitivity may be useful when process separation is reciprocal, but it does not make either member preferable to NCD in advance. Figure 12 shows that the result is stable across the family. Under the capability-plus-cross-control specification, the cross-benchmark means are -0.209 for residualised V_1, -0.212 for V_2, and -0.216 for V_∞. For every interpolant, the compression estimate is negative on all ten benchmarks, while the cross-controlled semantic estimate is positive on nine of ten. The compression estimates have Spearman agreement 0.988 across interpolants and a mean absolute difference of 0.019, compared with a mean

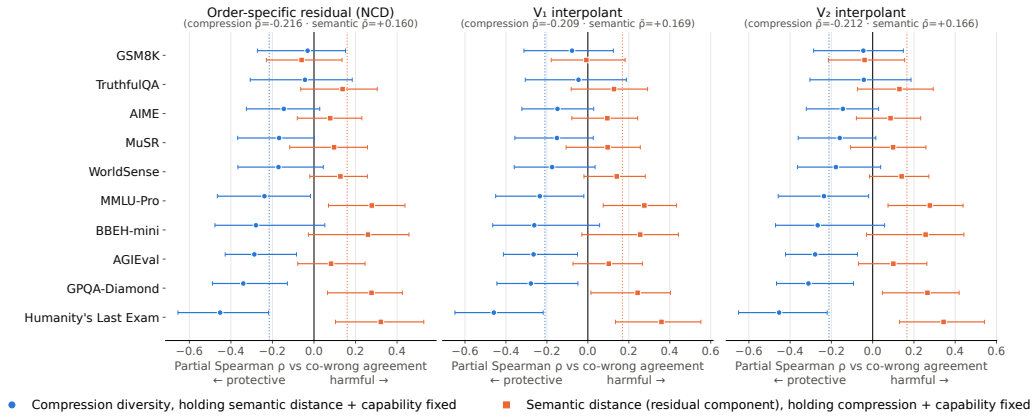

Figure 12: **The compression–failure association is stable across the V_p family.** Panels show the primary residualised V_∞ measure (NCD), V_1, and V_2. Blue circles are partial Spearman correlations between compression diversity and chance-corrected CWA, holding semantic distance and capability fixed; orange squares reverse the cross-control, estimating semantic distance while holding compression diversity and capability fixed. Horizontal bars are 95% model-node-bootstrap intervals, solid vertical lines mark zero, and dotted lines mark the cross-benchmark means reported above each panel. Compression estimates are negative on all ten benchmarks for every interpolant and vary little across panels relative to their uncertainty; semantic estimates are positive on nine of ten.

node-bootstrap interval width of 0.149. The small shifts between panels therefore support robustness across the family.

C Outcome Measures

C.1 Epoch-native CWA

Equation 3 pools question-level numerator and denominator counts into a single model-pair rate. Exactly-one-wrong response pairs contribute to the denominator, while the numerator records shared wrong answers. The epoch cross-product preserves repeated attraction to the same wrong answer through the response counts $c_i(a)c_j(a)$.

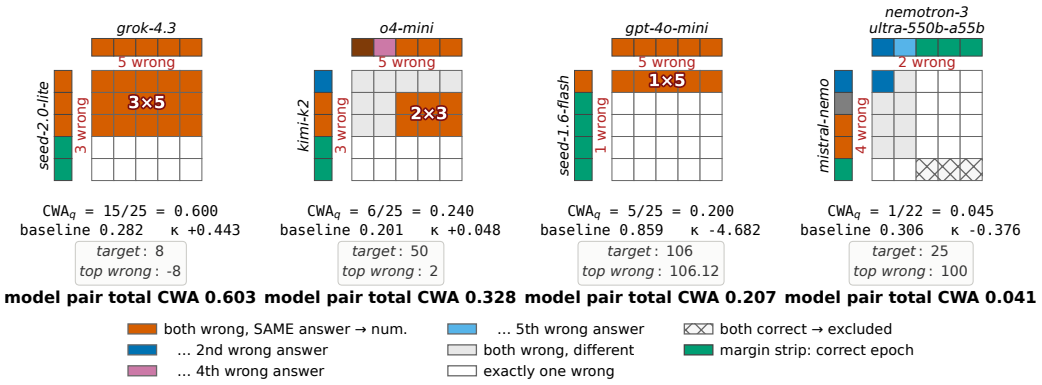

Figure 13: **Epoch-native CWA construction on GSM8K, complementing the GPQA-Diamond example in Figure 1.**

C.2 Chance baselines and relation to CAPA

Four baselines are computed. The per-question leave-pair-out baseline used in the primary analysis estimates each item's wrong-answer collision rate from the remaining models and then aggregates these expectations to the model-pair level. Pair-independence and pair-empirical baselines are comparators; a pooled-population baseline is retained only as a negative control because it does not condition on the question. This per-question-to-pair construction is used throughout the CWA analysis. Difference-form excess and chance-corrected κ rank pairs almost identically (Spearman 0.988 over benchmark means); eight of ten chance-correction denominators, $1 - \mathrm{CWA}^{\mathrm{exp}}$, lie in $[0.90, 0.98]$, with HLE-MC and MuSR the exceptions.

For completeness, consider a held-out pair (i, j). Let $b_{-ij,q}(a) = \sum_{k \notin \{i,j\}} c_{kq}(a)$ count wrong epochs from the remaining models that select $a \neq y_q$, and let $B_{-ij,q} = \sum_{a \neq y_q} b_{-ij,q}(a)$. On questions for which both held-out models have a wrong epoch and $B_{-ij,q} \geq 2$, the collision probability is

$$h_{ijq} = \frac{\sum_{a \neq y_q} b_{-ij,q}(a)[b_{-ij,q}(a) - 1]}{B_{-ij,q}(B_{-ij,q} - 1)}, \qquad \bar{h}_{-ij} = \frac{1}{|\mathcal{Q}_{ij}|} \sum_{q \in \mathcal{Q}_{ij}} h_{ijq}. \tag{23}$$

The baseline in Equation 4 multiplies $\bar{h}_{-ij}$ by $p_i p_j / (p_i + p_j - p_i p_j)$, where each p is the fraction of parsed epochs that are wrong across questions with parsed responses from both models.

CAPA applies the same algebraic chance correction but targets overall prediction agreement, counting matches whether models are correct or wrong (Goel et al., 2025). CWA instead conditions on at least one error and targets agreement on the same wrong answer. In this analysis, CAPA is computed from modal outputs using its uniform-distractor baseline, whereas CWA retains the epoch-native answer distribution and uses the question-specific leave-pair-out baseline. Figure 15 shows that the measures are related but not equivalent. Chance correction increases their rank agreement on seven of the eight fixed-option benchmarks, while HLE-MC moves in the opposite direction.

C.3 Correlated failure across empirical distance quantiles

Figure 14 bins pairs into fixed percentile bands within the evaluated model population. Semantic distance is effectively flat. Mean κ changes from 0.141 in the lowest band to 0.139 in the highest. The order-specific residual falls from 0.167 to 0.086, an observed 48% contrast between the endpoint bands, and remains downward-sloping in the 80–100th-percentile band. Capability-adjusted residuals show the same shape, declining 5.38 percentile points from the middle to the highest band.

The two other raw-byte V_p residuals reproduce the pattern. V_1 changes from 0.162 to 0.084 and V_2 from 0.163 to 0.085, also 48% endpoint contrasts. Thus the pattern across distance bands does not depend on one interpolant. Low residual-distance bands remain above the chance line, so low

inferred process diversity corresponds to excess correlated failure rather than merely the absence of a benefit. The range is population-relative. Raw NCD covers only $[0.48, 0.86]$ of its nominal $[0, 1]$ interval, and the upper band is not an absolute maximum of generative-process diversity.

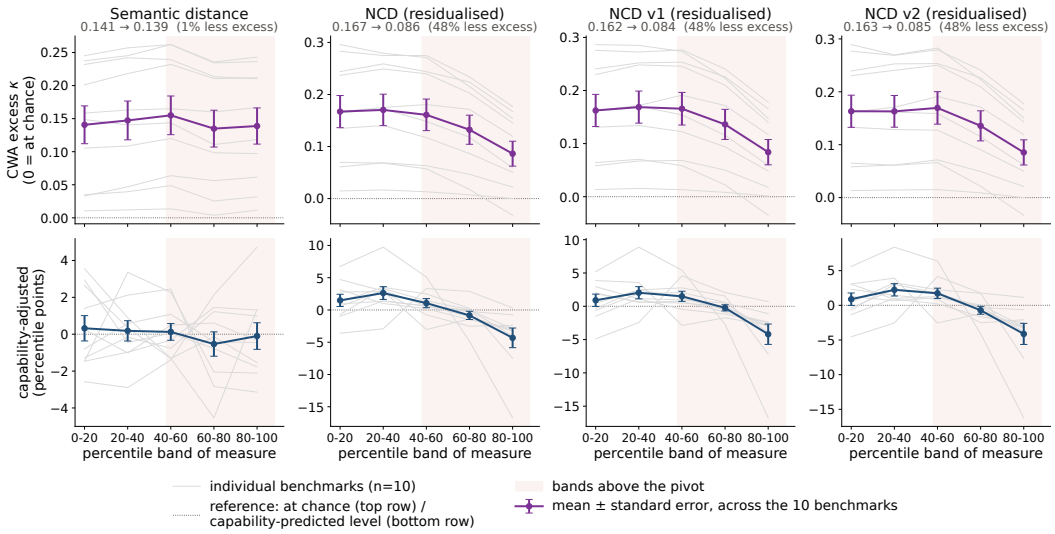

Figure 14: **Correlated failure across the range of each measure.** Pairs are binned by semantic distance, the primary order-specific NCD residual, and the raw-byte V_1 and V_2 residuals. Top: mean chance-corrected CWA; bottom: deviation from the capability-predicted level. Heavy lines show means and standard errors across ten benchmarks; light lines show individual benchmarks. Semantic distance is flat, whereas all three compression residuals decline by 48% from the lowest to highest band and remain downward-sloping in the observed upper tail.

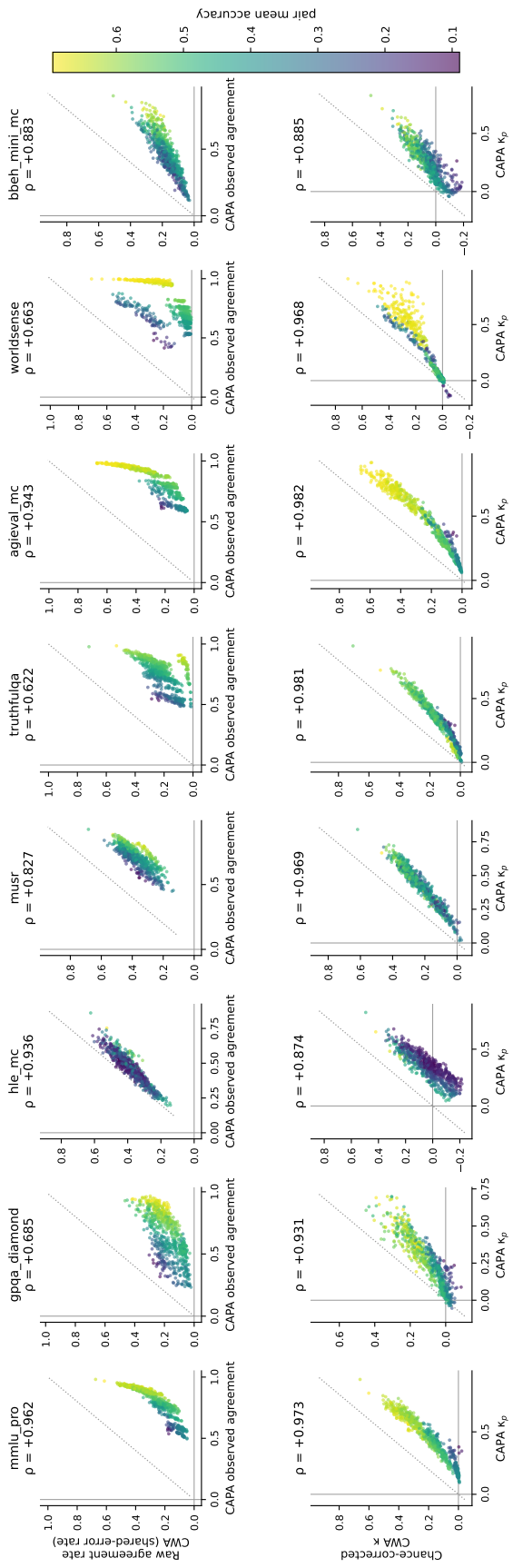

Figure 15: **CWA and CAPA measure related but distinct forms of agreement.** Each point is a model pair on one of eight fixed-option benchmarks; numeric-answer benchmarks are omitted because CAPA is undefined. The top row compares raw CAPA agreement with CWA's shared-error rate, and the bottom row compares their chance-corrected forms. Points are coloured by pair mean accuracy, dotted lines show $y = x$, and panel headings report Spearman ρ. Raw CWA generally lies below CAPA because it excludes both-correct agreement. Rank agreement spans 0.62–0.96 before correction and 0.87–0.98 after correction; HLE-MC is the exception, decreasing from 0.936 to 0.874.

D ENSEMBLE SELECTION

This appendix gives the definitions, kernel, matching procedure, and inference behind Section 2.6 and Figure 4, with enough detail to reproduce the analysis from the per-model benchmark records and the two 38×38 distance matrices. Numbers are for the primary residual V_∞ unless stated; interpolant results are tabulated in Appendix D.7. At the ensemble level the chance-corrected κ_S of Equation 5 is the primary outcome and the difference form $\mathrm{CFR}_S - \mathrm{CFR}_S^{\exp}$ is the companion, matching the pairwise convention.

D.1 ENUMERATION AND ELIGIBILITY

For $k = 5$, every one of the $\binom{38}{5} = 501{,}942$ ensembles is enumerated per benchmark key and per residual variant and scored on the epoch-native outcomes below. Two estimability rules define the eligible population, and both are applied to the selection universe and to every control pool. First, an ensemble's chance-corrected outcomes are undefined when fewer than 10 items are estimable for it. Second, only ensembles with at least 30 estimable items enter, the same threshold as the pairwise conditioning gate. A key is analysed when at least $10T = 2{,}000$ eligible ensembles remain, where $T = 200$ is the selection size. Table 2 gives the eligible counts, which are identical across the three variants because eligibility depends on outcomes and capability only. Eleven keys are analysed, collapsing to nine source benchmarks; the three BBEH-mini answer spaces pool into one source, question-weighted 180/120/160. GSM8K is excluded by this guard, not by choice, and remains in the ten-benchmark pairwise analysis.

Table 2: **Eligible five-model ensembles per benchmark key** out of 501,942 enumerated.

Key	Eligible	Key	Eligible
HLE	501,942	TruthfulQA	376,992
MMLU-Pro	501,942	BBEH-mini (numeric)	205,200
MuSR	501,941	GPQA-Diamond	97,063
AGIEval	501,280	AIME	16,813
BBEH-mini (other)	500,435	WorldSense	14,522
BBEH-mini (MC)	490,908	GSM8K	1,414 (excluded)

D.2 OUTCOME

On item q, with n_m parsed epochs for member m, let $T_q = \prod_m n_m$ be the number of epoch tuples, A_q the tuples in which every member is wrong, and C_q the tuples in which every member is wrong and all give the same wrong answer. The unconditional convergent-failure rate is $\sum_q C_q / \sum_q T_q$ and the conditional rate is $\mathrm{CFR}_S = \sum_q C_q / \sum_q A_q = P(\text{same} \mid \text{all wrong})$. The conditional form is primary for the same reason CWA is conditional at the pairwise level. It separates *how* an ensemble failed from *whether* it failed, so a diverse ensemble is not rewarded merely for being accurate.

On each item where every member has at least one wrong epoch, the leave-ensemble-out population is the pooled wrong epochs of the other $M - k$ models, N_q in total with n_x on wrong answer x. The chance that k draws without replacement all land on one wrong answer is the hypergeometric collision probability b_q of Equation 5. An item is estimable when every member has a wrong epoch, $N_q \geq 8$, and $\binom{N_q}{k} > 0$. With E the estimable set,

$$\mathrm{CFR}_S = \frac{\sum_{q \in E} C_q}{\sum_{q \in E} A_q}, \qquad \mathrm{CFR}_S^{\exp} = \frac{\sum_{q \in E} A_q\, b_q}{\sum_{q \in E} A_q}, \qquad \kappa_S = \frac{\mathrm{CFR}_S - \mathrm{CFR}_S^{\exp}}{1 - \mathrm{CFR}_S^{\exp}}. \tag{24}$$

Observed and expected terms share the item set and the all-wrong tuple weights, so the difference is zero whenever observed convergence equals chance item by item. This alignment is stronger than the pairwise baseline of Equation 4, whose $\bar{h}_{-ij}$ is an unweighted item mean. At one epoch per cell the construction reduces to the modal leave-ensemble-out baseline, and at $k = 2$ it reduces to the pairwise per-item collision rate $\sum_x c_x(c_x - 1)/[N(N-1)]$ of Appendix C.

D.3 KERNEL, VENDI SCORE, AND LOG-DETERMINANT

Let D be the 38×38 signed residual matrix d^{IGP}, or the non-negative semantic distance matrix $1 -$ similarity. With one bandwidth per measure fixed from the median absolute off-diagonal entry,

$$\gamma = \frac{1}{\mathrm{median}_{i<j} |D_{ij}| + 10^{-9}}, \tag{25}$$

each model i receives a two-sided decaying profile over all M models as landmarks and is normalised to unit length,

$$\phi_i = \left[e^{-\gamma \max(D_{ij},0)}\right]_{j=1}^{M} \,\|\, \left[e^{-\gamma \max(-D_{ij},0)}\right]_{j=1}^{M}, \qquad u_i = \frac{\phi_i}{\|\phi_i\|_2}, \qquad K_{ij} = u_i^\top u_j, \tag{26}$$

where $\|$ denotes concatenation. Positive and negative residuals occupy separate channels, so a pair that is anomalously *convergent* (negative residual; 329 of 703 pairs) is not mapped to a large distance, which is the defect of a squared-distance Gaussian kernel on a signed matrix. K is positive semidefinite with unit diagonal by construction. It is a chosen geometry for signed residuals, not a claim that the residual matrix is itself a similarity. On non-negative semantic distance the negative channel is a constant block and the same construction is used. Because the pool is the full fleet on every key, the kernels are identical across keys: $\gamma_{\mathrm{IGP}} = 89.46$ (median $|D| \approx 0.011$) and $\gamma_{\mathrm{sem}} = 3.23$ (median ≈ 0.31).

For a candidate S, take the 5×5 principal submatrix K_S with eigenvalues $\lambda_1, \ldots, \lambda_5 \geq 0$; round-off negatives are clipped and a materially indefinite submatrix is treated as an error. Then

$$\mathrm{logdet}(S) = \sum_r \log(\lambda_r + 10^{-9}), \qquad p_r = \frac{\lambda_r}{\sum_s \lambda_s}, \qquad \mathrm{Vendi}(S) = \exp\Big(-\sum_r p_r \log p_r\Big), \tag{27}$$

with $0 \log 0 = 0$. Vendi is the exponential of the von Neumann entropy of the trace-normalised kernel, the effective rank of the ensemble in feature space (Friedman & Dieng, 2023), and lies in $[1, 5]$; it receives no ridge. The log-determinant is the determinantal-point-process volume objective (Kulesza & Taskar, 2012) and is evaluated as an alternative selector only. Nothing is recentred or refitted per candidate, and no outcome enters any score. Every eligible ensemble is scored by exact batched eigendecomposition, and selection is the exact top $T = 200$ with ties broken in enumeration order.

Three measured properties qualify the reading of Vendi. First, it is not a monotone function of mean distance: Spearman agreement with mean pairwise distance is 0.37–0.48 for compression and 0.22–0.43 for semantic distance across the 33 key-by-variant pools. Second, on this unit-diagonal kernel, uncentred Vendi is rank-equivalent to the trace of the centred submatrix (Spearman 0.98 for compression, 1.00 for semantic). It is therefore a *magnitude* score of how spread the members are, not a score of how evenly that spread is distributed across directions, and we do not attach volume or independent-mode language to it. Third, the semantic kernel is close to rank one under this bandwidth, so semantic Vendi has a narrow range (top-200 cut-offs 1.010–1.020 against control means 1.009–1.014, where compression cut-offs are 1.62–2.10 against 1.52–1.84). The semantic ranking remains well defined, since semantic-Vendi selections reach the 87th–95th percentile of mean semantic distance, and the mean-distance selector, which has no such compression of range, gives the same sign pattern under the pin design (Table 3).

D.4 ARMS AND PIN BANDS

For each key, variant, and selector we run four arms. Two are unconstrained: the top 200 by the compression score and the top 200 by the semantic score. Two are pinned: the top 200 by the compression score among ensembles whose *mean semantic distance* lies in a band, and the top 200 by the semantic score among ensembles whose *mean compression distance* lies in a band. Bands are $[0, 20)$, $[20, 40)$, $[40, 60)$, $[60, 80)$, and $[80, 100]$ in population percentiles of the eligible pool's mean distance on the held measure. The pin is always on mean distance, also for spectral selectors, so the held quantity and the dose annotations in Figure 4 (median percentile of the free measure's mean distance among selected ensembles) share one scale. A band with fewer than 2,000 candidates is skipped.

D.5 CAPABILITY-MATCHED CONTROLS

Matching coordinates are an ensemble's mean member accuracy and mean pairwise absolute accuracy gap, each divided by its SD over the eligible pool. A control matches when it lies within a Chebyshev window of $\pm\tau$ SD on both coordinates; τ starts at 0.5, grows by a factor of 1.5 until at least 50 controls match, and is capped at 2.0. The gap is matched because level-only matching was measured to book capability spread as diversity benefit. For pinned arms two hard constraints precede the capability window: a control must lie in the same band, and its held mean distance must lie within 0.1 population SD of the selected ensemble's value. The selected ensemble satisfies its own constraints and is therefore a member of its own pool (the matched-*random* definition); a pool containing only the selection contributes exactly zero and is flagged. All 200 selections are retained with whatever non-empty pool they have, and each cell reports pool minimum, median, and maximum, the count below target, and the count of self-only pools. Held-axis imbalance is at most 0.008 SD in absolute value across all pinned Vendi cells, and capability imbalance has median 0.06–0.08 SD and maximum 0.19 SD. Vendi pool medians per selection range from 79 (AIME, 0–20 band) to 6,243 (AGIEval, 60–80 band).

D.6 ESTIMAND AND INFERENCE

For one cell, with y_s the outcome of selection s and $\bar{y}_{\mathcal{C}(s)}$ the mean over its matched pool,

$$\Delta = \frac{1}{T}\sum_{s=1}^{T} y_s - \frac{1}{T}\sum_{s=1}^{T} \bar{y}_{\mathcal{C}(s)}. \tag{28}$$

Negative Δ means the selected ensembles show less chance-corrected agreement, given that all five members are wrong, than capability-matched ensembles at the same held distance. Within a cell the interval is a paired bootstrap over selections with 1,000 resamples: resample the 200 selections with replacement and, for each, draw one control from *its own* pool, so the pairing travels with the selection; the 2.5th and 97.5th percentiles of the resampled Δ are reported, and both arms are resampled. This interval is *selection-level*. Top-200 arms are built from roughly twenty distinct models, so it does not reproduce model-level uncertainty and is read as descriptive precision. Across benchmarks, per-key Δ are collapsed within source (BBEH-mini parts by item count), averaged with one vote per source, and reported as the mean with a 95% t-interval over the nine sources ($t_{0.975,8} = 2.31$) together with the median. Selector families are never pooled, no benchmark is excluded on the basis of its result, and no multiplicity correction is applied across the 90 pinned cells per outcome (5 bands, 3 variants, 2 arms, 3 selectors).

D.7 ADDITIONAL RESULTS

Table 4 the difference-form companion, and Table 3 the three selector families on the V_∞ residual. The qualitative pattern, compression-pinned arms negative and semantic-pinned arms non-negative, holds across all three selector families. The unconstrained arms do not separate the measures. Under mean distance the unconstrained semantic arm is also below matched random on every source, as collinearity predicts, and under the pin design the semantic arm is significantly *positive* in the lowest compression band for mean distance and in three mid bands for log-determinant. Per-source detail for the Vendi compression arm in the 60–80 band on the difference scale is HLE -0.036, AGIEval -0.014, GPQA-Diamond -0.009, WorldSense -0.009, MMLU-Pro -0.007, TruthfulQA -0.006, BBEH-mini -0.003, MuSR -0.002, and AIME $+0.002$; AIME's difference scale is an order of magnitude smaller than the others'. The unconstrained compression arm is dominated by HLE (-0.077); without it the mean over the remaining eight sources is about -0.010. Figures 16–18 show the Vendi, log-determinant, and mean-distance selectors in the layout of Figure 4, for the V_∞ residual and the interpolants V_1 and V_2.

Table 3: **Selector families on the V_∞ residual, difference-form outcome.** Mean over nine sources with 95% t-interval; the trailing count is the number of sources with negative Δ. Bold intervals exclude zero.

Arm	Band	Mean pairwise distance	Kernel log-det	Vendi
Compression, unconstrained	—	-0.027 $[-0.054, +0.000]$ 7	-0.015 $[-0.033, +0.004]$ 7	-0.018 $[-0.036, +0.001]$ 7
Semantic, unconstrained	—	$\mathbf{-0.025}$ $[\mathbf{-0.040}, \mathbf{-0.009}]$ 9	-0.001 $[-0.015, +0.013]$ 5	-0.005 $[-0.018, +0.007]$ 7
Compression, semantic held	0–20	-0.006 $[-0.022, +0.010]$ 6	-0.009 $[-0.026, +0.008]$ 6	-0.010 $[-0.026, +0.005]$ 7
	20–40	-0.009 $[-0.028, +0.009]$ 4	-0.011 $[-0.030, +0.008]$ 5	-0.014 $[-0.032, +0.004]$ 7
	40–60	-0.013 $[-0.036, +0.010]$ 5	-0.012 $[-0.028, +0.003]$ 6	-0.013 $[-0.027, +0.001]$ 7
	60–80	-0.019 $[-0.046, +0.007]$ 7	-0.009 $[-0.020, +0.001]$ 6	$\mathbf{-0.010}$ $[\mathbf{-0.018}, \mathbf{-0.001}]$ 8
	80–100	-0.017 $[-0.039, +0.004]$ 7	-0.009 $[-0.021, +0.003]$ 7	-0.013 $[-0.029, +0.004]$ 7
Semantic, compression held	0–20	$\mathbf{+0.024}$ $[\mathbf{+0.002}, \mathbf{+0.046}]$ 1	$+0.012$ $[-0.001, +0.024]$ 2	$+0.003$ $[-0.005, +0.011]$ 2
	20–40	$+0.019$ $[-0.004, +0.043]$ 1	$\mathbf{+0.008}$ $[\mathbf{+0.001}, \mathbf{+0.015}]$ 2	-0.000 $[-0.010, +0.010]$ 4
	40–60	$+0.010$ $[-0.008, +0.029]$ 1	$\mathbf{+0.011}$ $[\mathbf{+0.002}, \mathbf{+0.020}]$ 2	$+0.005$ $[-0.008, +0.018]$ 3
	60–80	$+0.002$ $[-0.012, +0.017]$ 4	$\mathbf{+0.010}$ $[\mathbf{+0.000}, \mathbf{+0.020}]$ 2	$+0.004$ $[-0.009, +0.017]$ 4
	80–100	-0.004 $[-0.015, +0.006]$ 7	$+0.008$ $[-0.006, +0.021]$ 4	$+0.007$ $[-0.005, +0.019]$ 4

Table 4: **Vendi selector, difference-form companion $\mathrm{CFR}_S - \mathrm{CFR}_S^{\mathrm{exp}}$.** † marks at least one selection with fewer than 50 controls.

Arm	Band	V_∞ residual	V_1 residual	V_2 residual
Compression Vendi, semantic held	0–20†	-0.010 $[-0.026, +0.005]$	-0.013 $[-0.033, +0.008]$	-0.011 $[-0.029, +0.006]$
	20–40	-0.014 $[-0.032, +0.004]$	-0.014 $[-0.033, +0.006]$	-0.014 $[-0.034, +0.005]$
	40–60	-0.013 $[-0.027, +0.001]$	-0.012 $[-0.029, +0.005]$	-0.012 $[-0.028, +0.003]$
	60–80	$\mathbf{-0.010}$ $[\mathbf{-0.018}, \mathbf{-0.001}]$	-0.009 $[-0.020, +0.001]$	$\mathbf{-0.010}$ $[\mathbf{-0.019}, \mathbf{-0.0003}]$
	80–100	-0.013 $[-0.029, +0.004]$	-0.014 $[-0.030, +0.003]$	-0.013 $[-0.030, +0.004]$
Semantic Vendi, compression held	0–20†	$+0.003$ $[-0.005, +0.011]$	$+0.003$ $[-0.005, +0.011]$	$+0.003$ $[-0.005, +0.011]$
	20–40	-0.000 $[-0.010, +0.010]$	-0.002 $[-0.011, +0.008]$	$+0.001$ $[-0.009, +0.010]$
	40–60	$+0.005$ $[-0.008, +0.018]$	$+0.003$ $[-0.009, +0.016]$	$+0.005$ $[-0.008, +0.017]$
	60–80	$+0.004$ $[-0.009, +0.017]$	$+0.004$ $[-0.009, +0.017]$	$+0.005$ $[-0.009, +0.018]$
	80–100	$+0.007$ $[-0.005, +0.019]$	$+0.008$ $[-0.004, +0.021]$†	$+0.008$ $[-0.005, +0.020]$†
Compression Vendi, unconstrained	—	-0.018 $[-0.036, +0.001]$	-0.014 $[-0.032, +0.004]$	-0.015 $[-0.034, +0.003]$
Semantic Vendi, unconstrained	—	-0.005 $[-0.018, +0.007]$	-0.005 $[-0.018, +0.007]$	-0.005 $[-0.018, +0.007]$

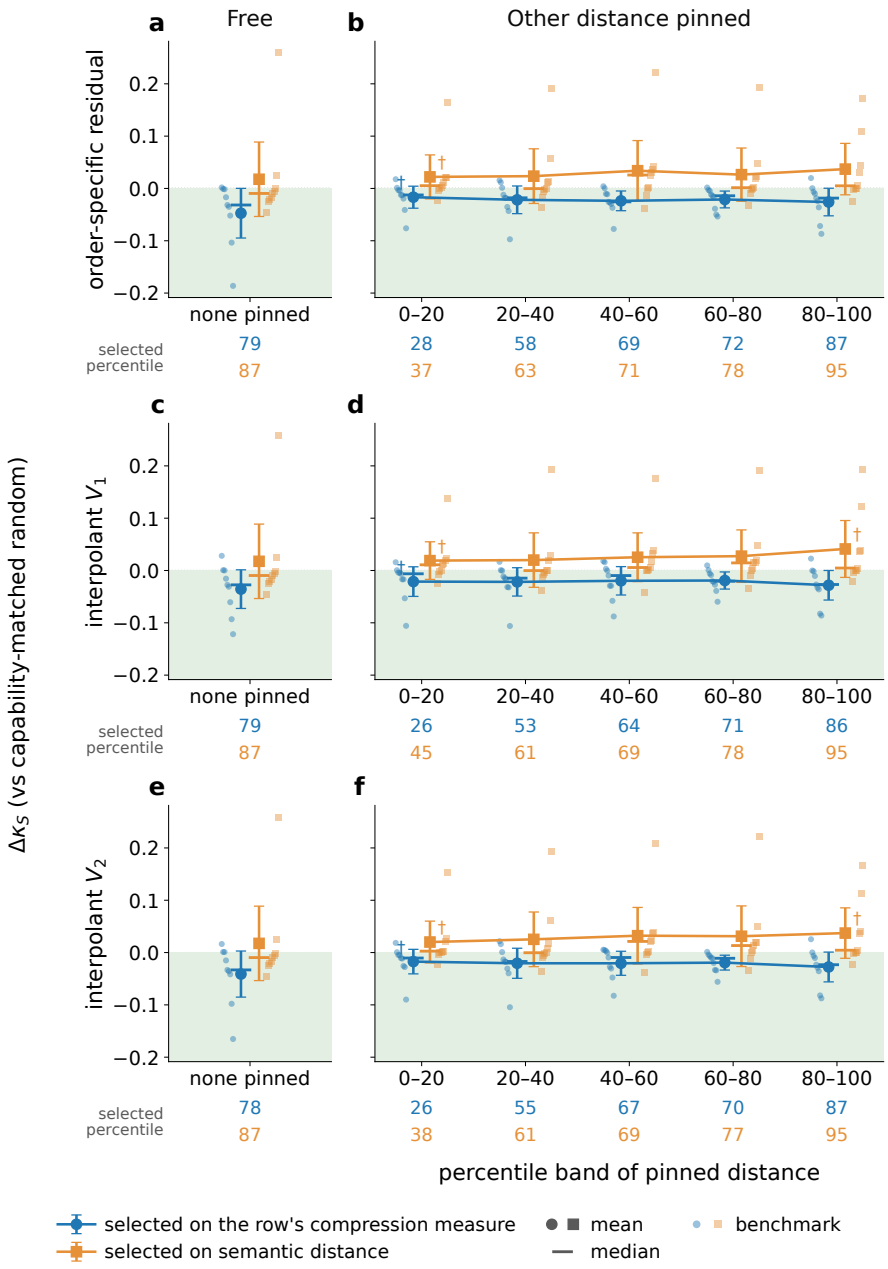

Figure 16: **Vendi selection free and pinned to a percentile band of the other distance, for each compression measure.** As Figure 4, whose panels are (a, b) here. Rows are the order-specific residual (a, b) and its interpolants V_1 (c, d) and V_2 (e, f). Large markers are the difference in κ_S (Equation 5) between the top 200 five-model ensembles by Vendi score of one measure and capability-matched random ensembles, averaged over nine source benchmarks; small points are the individual benchmarks. Left column: free selection. Right column: the other measure's mean distance pinned to each percentile band. Bars are 95% t-intervals and horizontal marks are medians; the green region below zero marks lower κ_S than capability-matched random. Controls match mean accuracy and accuracy gap, and in the right column lie in the same band within 0.1 SD of the pinned distance. Coloured numerals are the median percentile of the selected ensembles on the selected measure. Daggers mark bands in which at least one selection had fewer than 50 controls.

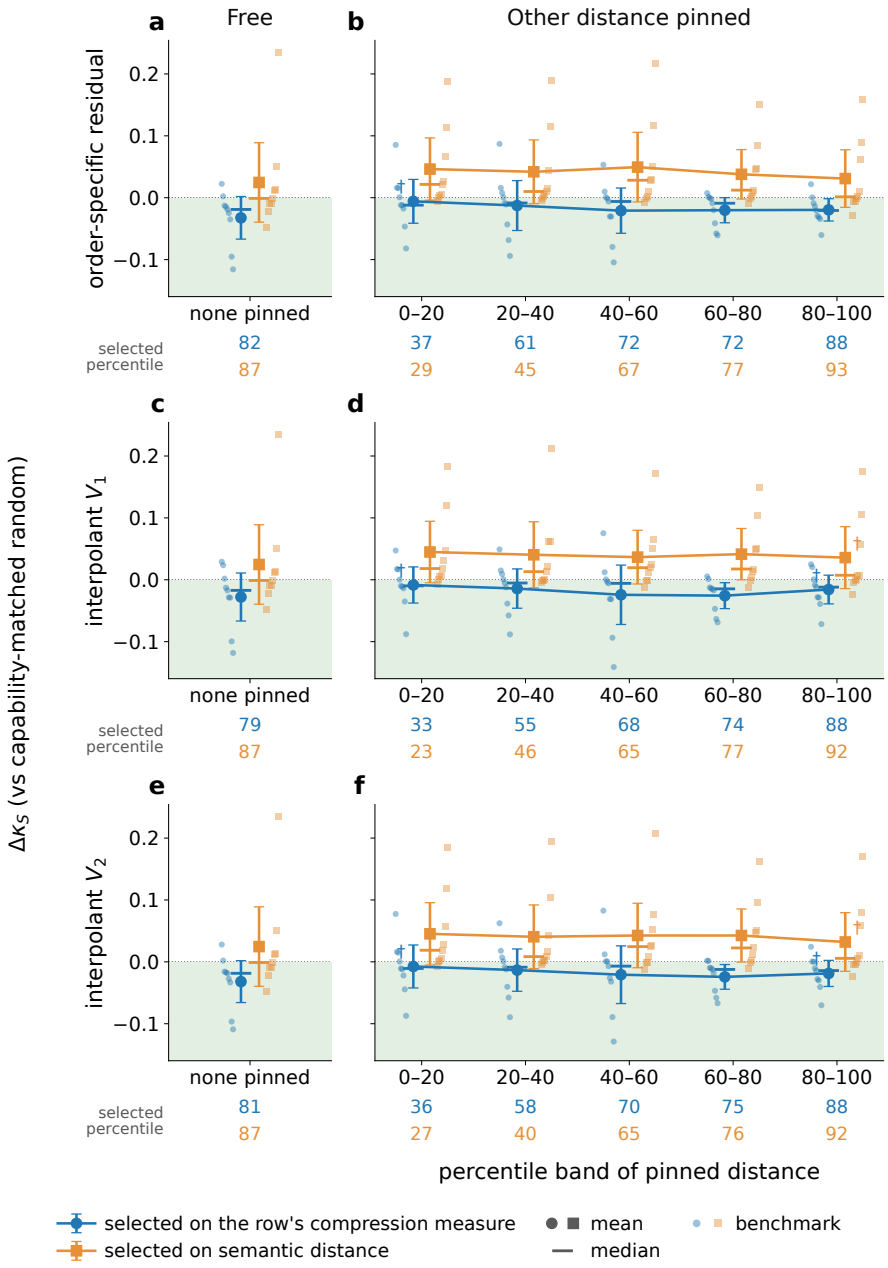

Figure 17: **Log-determinant selection free and pinned to a percentile band of the other distance, for each compression measure.** As Figure 4, with ensembles selected by the log-determinant of the distance kernel. Rows are the order-specific residual (a, b) and its interpolants V_1 (c, d) and V_2 (e, f). Large markers are the difference in κ_S (Equation 5) between the top 200 five-model ensembles by log-determinant of one measure and capability-matched random ensembles, averaged over nine source benchmarks; small points are the individual benchmarks. Left column: free selection. Right column: the other measure's mean distance pinned to each percentile band. Bars are 95% t-intervals and horizontal marks are medians; the green region below zero marks lower κ_S than capability-matched random. Controls match mean accuracy and accuracy gap, and in the right column lie in the same band within 0.1 SD of the pinned distance. Coloured numerals are the median percentile of the selected ensembles on the selected measure. Daggers mark bands in which at least one selection had fewer than 50 controls.

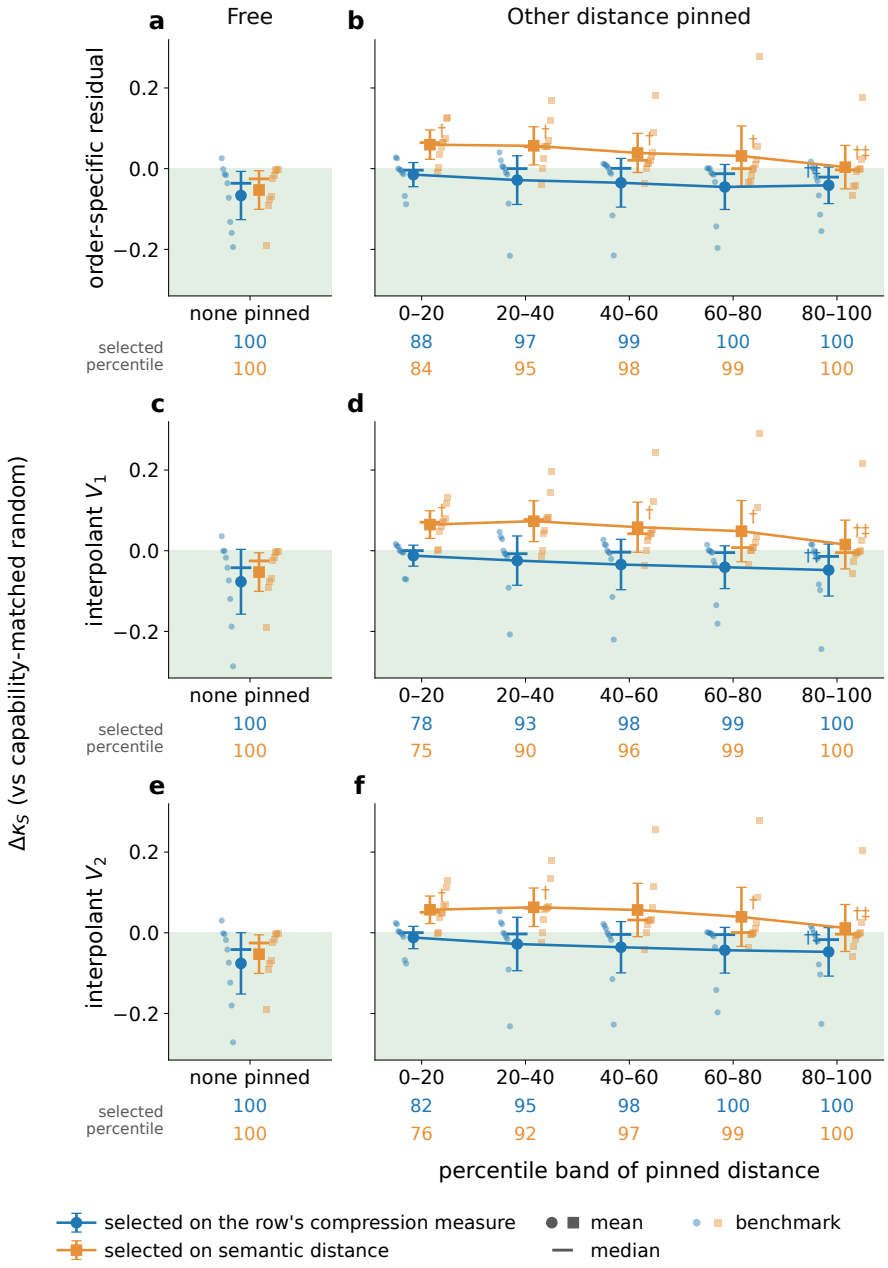

Figure 18: **Mean-distance selection free and pinned to a percentile band of the other distance, for each compression measure.** As Figure 4, with ensembles selected by mean pairwise distance. Rows are the order-specific residual (a, b) and its interpolants V_1 (c, d) and V_2 (e, f). Large markers are the difference in κ_S (Equation 5) between the top 200 five-model ensembles by mean pairwise distance of one measure and capability-matched random ensembles, averaged over nine source benchmarks; small points are the individual benchmarks. Left column: free selection. Right column: the other measure's mean distance pinned to each percentile band. Bars are 95% t-intervals and horizontal marks are medians; the green region below zero marks lower κ_S than capability-matched random. Controls match mean accuracy and accuracy gap, and in the right column lie in the same band within 0.1 SD of the pinned distance. Coloured numerals are the median percentile of the selected ensembles on the selected measure. Daggers mark bands in which at least one selection had fewer than 50 controls; double daggers mark bands in which at least one selection's only control was itself.

E STATISTICAL METHODS

E.1 ESTIMAND AND PARTIAL RANK CORRELATION

For paired vectors x and y, Pearson's correlation is the covariance standardised by their sample standard deviations,

$$r(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}. \tag{29}$$

It measures linear association on the observed scale. Spearman's ρ_s applies the same calculation to the componentwise midranks, $\rho_s(x,y) = r\{R(x), R(y)\}$, and therefore measures monotone association while being invariant to strictly increasing transformations (Spearman, 1904). We use Pearson correlations for explicitly linear-scale diagnostics and Spearman correlations when comparing pair or model orderings.

The primary estimand is the partial Spearman association between a pair's compression residual and chance-corrected CWA, conditional on semantic distance, capability level, and capability gap. Let $\widetilde{x} = R(x)$ and $\widetilde{y} = R(y)$, and let $\widetilde{Z}$ contain an intercept and the midranks of all controls. With $P_Z = \widetilde{Z}(\widetilde{Z}^\mathsf{T}\widetilde{Z})^{-1}\widetilde{Z}^\mathsf{T}$, we compute

$$\rho_s(x, y \mid Z) = r\left\{(I - P_Z)\widetilde{x}, (I - P_Z)\widetilde{y}\right\}. \tag{30}$$

Thus, the ranked predictor and outcome are residualised separately against the same ranked controls and the two residual vectors are correlated. Simultaneous control matters. Holding capability gap alone gives -0.037, whereas holding level and gap gives -0.148. Rank partialling removes dependence linear in ranks and can absorb monotone nonlinear confounding, but can leave residue from confounding additive in raw values.

Capability level is $(a_i + a_j)/2$ and gap is $|a_i - a_j|$. The gap enters as a pre-existing pair attribute. Because capability is estimated from the same benchmark correctness matrices, both terms function as analytic controls for pair-level performance. Measured estimates are -0.026 under raw-linear capability control, -0.110 under quadratic control, and -0.143 under rank control. The empirical relationship supports the rank specification used in the primary analysis.

E.2 DYADIC DEPENDENCE AND MODEL-NODE RESAMPLING

To make the dependence explicit, write a generic pair quantity as $x_{ij} = \mu + a_i + a_j + e_{ij}$, where the independent node effects have variance $\mathrm{Var}(a_i) = \sigma_a^2$, the independent dyad residuals have variance $\mathrm{Var}(e_{ij}) = \sigma_e^2$, and $\bar{x}$ averages all unordered pairs. Then

$$\mathrm{Var}(\bar{x}) = \frac{4\sigma_a^2}{M} + \frac{\sigma_e^2}{\binom{M}{2}}. \tag{31}$$

The model component decreases with the number of nodes M, not with the number of dyads $\binom{M}{2}$, because each a_i is shared by the $M - 1$ pairs incident to model i. This component is material in the observed data. An incidence model attributes 58.6% of compression distance's rank variance and 61–85% of outcome rank variance to model-level structure.

The bootstrap preserves this incidence structure directly. For each replicate, we draw M model slots with replacement, construct every unordered pair of distinct slots, map those induced dyads to the observed pair rows, and recompute the complete rank-partial statistic, including all controls. Selecting a model more than once repeats all of its incident dyads; self-pairs and non-estimable dyads are omitted. The interval is given by the 2.5th and 97.5th percentiles of the resulting statistic. A delete-one-model jackknife closely agrees with these intervals on every benchmark (mean width ratio 0.95, range 0.89–1.02). In 300 simulations in which the true association was zero, calibrated to the measured model-level variance shares, the node-bootstrap interval excluded zero in 3.0% of simulations at a nominal 5% level. The intervals remain conditional on prompts, generations, benchmark questions, evaluation epochs, and the composition of the evaluated model population.

E.3 PAIRED COMPARISON OF CONDITIONAL ASSOCIATIONS

The two cross-controlled partial correlations are estimated on the same model pairs and outcome within each benchmark. Let Z_b contain capability level and gap for benchmark b, and define

$$\begin{aligned} \rho_b^{\mathrm{IGP}} &= \rho_s(d^{\mathrm{IGP}}, \kappa_b \mid d^{\mathrm{sem}}, Z_b), \\ \rho_b^{\mathrm{sem}} &= \rho_s(d^{\mathrm{sem}}, \kappa_b \mid d^{\mathrm{IGP}}, Z_b), \qquad \Delta_b = \rho_b^{\mathrm{IGP}} - \rho_b^{\mathrm{sem}}. \end{aligned} \tag{32}$$

Both coefficients are partial correlations on ranked variables and therefore lie in $[-1, 1]$. Their difference Δ_b is expressed in correlation units but is not itself a correlation coefficient; its theoretical range is $[-2, 2]$. Negative values indicate that the order-specific compression residual has the more negative association with correlated failure.

The two coefficients are dependent because they share the same models, dyads, outcome, and control variables. We therefore compute both coefficients within each model-node bootstrap replicate and difference them within that replicate. The percentile interval for each benchmark consequently retains the sampling covariance between the two estimates. Comparing the overlap of their marginal intervals would not test $\Delta_b = 0$. Across benchmarks, we report the unweighted mean $\bar{\Delta} = 10^{-1} \sum_b \Delta_b$ and its t-interval over the ten benchmark estimates. Figure 19 shows $\bar{\Delta} = -0.375$ with 95% interval $[-0.545, -0.206]$.

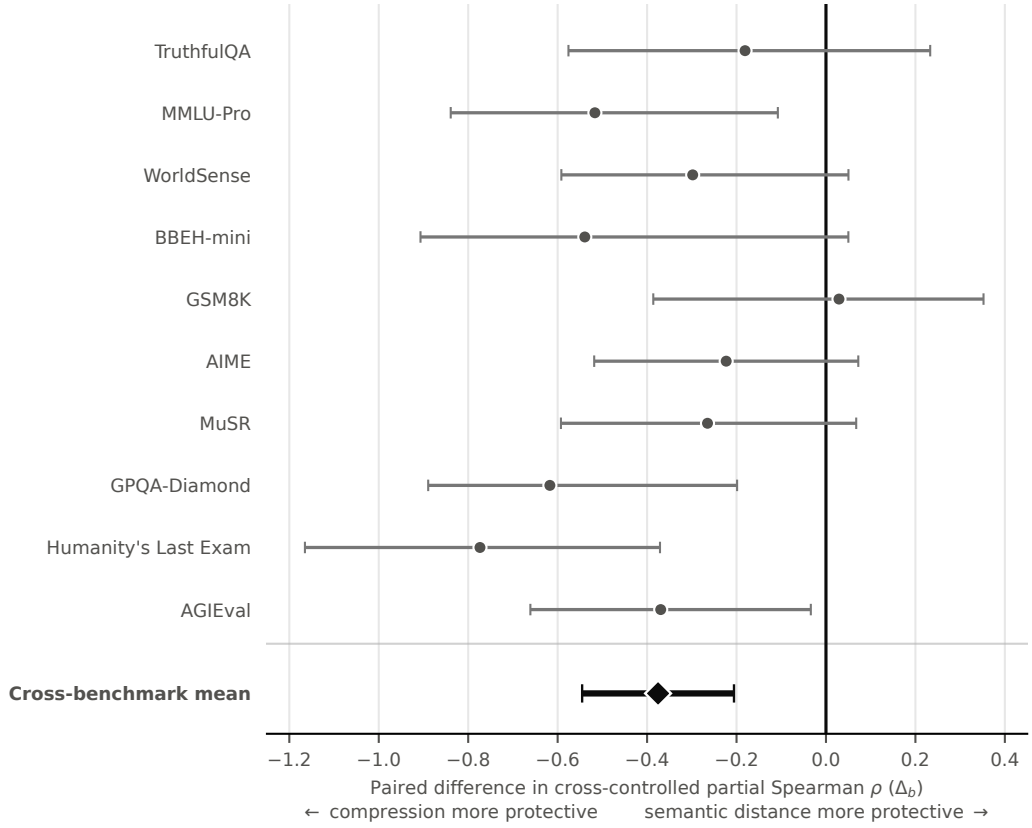

Figure 19: **Paired comparison of the cross-controlled partial correlations.** Each benchmark row shows $\Delta_b = \rho_b^{\mathrm{IGP}} - \rho_b^{\mathrm{sem}}$, a difference in correlation units with theoretical range $[-2, 2]$. Circles are point estimates and horizontal bars are 95% model-node-bootstrap intervals obtained by differencing the two correlations within each resample. The diamond is the unweighted mean across ten benchmarks and its bar is the corresponding 95% t-interval. Negative values indicate that the order-specific compression residual has the more negative conditional association with correlated failure.

E.4 Repeated-measures ranks and multiplicity

The within-model analysis compares $K = 38$ models repeatedly across $B = 100$ prompt blocks. Within each prompt, the models are ranked on the response-variability measure. Let R_j be the rank sum for model j across prompts and $\bar{R} = B(K+1)/2$. In the absence of ties, Kendall's coefficient of concordance is

$$W = \frac{12\sum_{j=1}^{K}(R_j - \bar{R})^2}{B^2(K^3 - K)}, \qquad 0 \le W \le 1, \tag{33}$$

with the standard tie correction used when ranks coincide (Kendall & Babington Smith, 1939). Here, $W = 0$ indicates no stable model ordering across prompts and $W = 1$ indicates complete agreement among the prompt-specific rankings. It is an effect-size measure, not a test of any particular model pair.

The Friedman test uses the same blocked ranks to test the omnibus hypothesis that the models have no systematic differences in rank location across prompts (Friedman, 1937). Its tie-corrected statistic Q satisfies $Q = B(K-1)W$ and is compared with a χ^2_{K-1} reference distribution. Rejection establishes that at least one model differs in its repeated rank pattern; it neither identifies which models differ nor supplies pairwise comparisons. The W values in Appendix B.6 describe the magnitude of the stable ordering, while the associated Friedman tests assess whether that ordering is distinguishable from the equal-rank hypothesis.

The compressor analysis uses a different paired procedure. For each of the $U = 72$ matched length–replicate units, let d_u be PPMd's shared-context effect minus that of one alternative compressor. The one-sided sign-flip test compares the observed $\bar{d}$ with 20,000 values $\bar{d}^*_b = U^{-1}\sum_u s_{bu} d_u$, where the signs $s_{bu} \in \{-1, +1\}$ are sampled independently with equal probability. Its Monte Carlo value is

$$p = \frac{1 + \#\{b : \bar{d}^*_b \ge \bar{d}\}}{20{,}001}, \tag{34}$$

so a reported value cannot be zero. The three PPMd-versus-compressor tests form one comparison family. If their ordered unadjusted values are $p_{(1)} \le \cdots \le p_{(m)}$, with $m = 3$, the Holm-adjusted values are

$$\widetilde{p}_{(i)} = \min\left\{1, \max_{1 \le j \le i}(m - j + 1)p_{(j)}\right\}. \tag{35}$$

This step-down adjustment controls the family-wise probability of at least one false rejection while retaining more power than applying the same Bonferroni threshold to every test (Holm, 1979). The value 1.5×10^{-4} reported in Appendix B.1 is the adjusted value for each of the three contrasts.

F Extended Related Work

F.1 Diversity and resilience in collective systems

Diversity can contribute to collective performance through several distinct mechanisms, including the buffering or insurance effects produced when components respond differently to fluctuating conditions (Holling, 1973; Yachi & Loreau, 1999). Functional diversity describes variation in the functions performed by system components. Response diversity describes variation in how components contributing to the same function respond to perturbation (Elmqvist et al., 2003; Ross et al., 2023). The latter is particularly relevant to redundancy. Components are not interchangeable safeguards if they respond identically under the conditions that cause failure. Work on resilience and critical transitions also shows that the organisation of heterogeneity and coupling can determine whether local perturbations remain local or become system-wide (Scheffer et al., 2012). Diversity is therefore multidimensional. Its effect depends on the system function and the disturbance under study.

Collective-adaptation research reaches a related conclusion from a problem-solving perspective. Heterogeneous information and strategies can expand the set of solutions explored by a group. Rapid information sharing can improve diffusion, but it can also accelerate convergence on a common and potentially suboptimal solution (Galesic et al., 2023; Tieman et al., 2025). The arrangement of diversity matters as much as its amount because coupling determines whether differences remain

available to the collective when conditions change. These findings motivate population-level analysis of AI systems, but they do not specify which representation of model diversity is appropriate. Our use of resilience theory is methodological rather than analogical. It determines the outcome-relative definition of diversity and the criterion used to validate its measure.

F.2 Diversity and correlated failure in AI systems

Multi-agent safety work has begun to identify failure modes that are not reducible to the behaviour of an isolated model. Homogeneous agents can respond synchronously to common information, propagate shared errors through a network, or provide ineffective mutual oversight (Hammond et al., 2025; Tomašev et al., 2026; Franklin et al., 2026). These risks extend concerns about algorithmic monoculture and are especially important for defence-in-depth designs whose reliability assumes that separate components fail under different conditions (Kleinberg & Raghavan, 2021).

Empirical evidence suggests that organisational and architectural labels are weak proxies for this independence. Language models from different providers select identical wrong answers at rates above pair-specific chance expectations (Goel et al., 2025; Kim et al., 2025). Provider count, parameter count, and model-family count may therefore overstate the effective diversity of a multi-model system. A useful audit must estimate behavioural relationships among the models themselves and validate those relationships against a relevant failure outcome.

The Artificial Hivemind measures one such relationship using embedding similarity among open-ended responses (Jiang et al., 2025). Its finding of high semantic similarity across the model population motivates our study and supplies the semantic baseline. Our objective differs in the property being estimated. Embedding similarity asks whether response meanings are similar. Inferred generative-process similarity asks whether observable sequences exhibit shared generative organisation. We evaluate whether this distinction explains variation in correlated failure that semantic similarity does not.

F.3 AIT and generative structure

AIT characterises an object by the length of the shortest program that generates it. Kolmogorov complexity is uncomputable, but it motivates universal similarity measures based on shared algorithmic information. Normalised Information Distance compares the information needed to describe either object given the other. NCD approximates this relation by replacing program length with compressed length (Cilibrasi & Vitányi, 2005; Vitányi et al., 2009). NCD has the practical advantage of operating directly on discrete sequences and requiring neither task-specific features nor a learned representation.

AIT approaches have also been used to distinguish generative regularity from apparent statistical randomness (Zenil et al., 2017; 2023). This perspective clarifies why generative and statistical descriptions need not coincide. An object may appear complex under one selected representation while retaining simple generative organisation. We make a narrower empirical claim than recovering a model's generating program. A compressor does not identify a latent algorithm, and finite-sample NCD is compressor-dependent. We use NCD as a comparative statistic of shared sequential structure and test its validity through prediction of disjoint correlated-failure outcomes.