

Supplementary Information

The evolution of sex for artificial intelligence:
a population-genetic framework for multigenerational model populations

Giorgio F. Gilestro

Department of Life Sciences, Imperial College London
giorgio@gilest.ro

Draft

Contents

SI Text S1–S4, SI Tables S1–S2, SI Methods M1–M7, SI Statistics, SI Figures S1–S16, and a separate Appendix 1, *The figures explained* (`figure_legends_for_students.pdf`), which restates every main and supplementary figure with a plain-language account of the experiment behind it, for readers from biology.

Reproducibility

Every experiment in this paper is defined by one committed configuration file under `configs/`. Running it produces three artifacts under `results/<name>/`: the results table (`results.parquet`), the fully resolved configuration, and a manifest recording content hashes, the master seed, and the git commit. Each experiment directory also contains a README with the figure legend and the current status of the experiment’s falsifier — the outcome that would refute its claim (see Methods M1) — plus a figure that regenerates from the parquet file alone. The script `reproduce.sh` re-runs the entire study from the master seeds, and `REPRODUCING.md` maps every panel of the manuscript to the configuration and seed behind it.

SI Text S1. The incompatibility floor: what no alignment can remove

Setting. Two models, A and B, are trained on the same input distribution. Their label functions f_A and f_B agree everywhere except on a *conflict set* S , whose size is its probability mass $\mu(S)$. In the conflict condition of the trained-network speciation experiment, S consists of the cyclically relabelled classes, so $\mu(S)$ is approximately the configured conflict fraction, up to class-balance corrections.

A *function-preserving transformation* T is any change to a network’s weights that leaves its outputs untouched. For a plain ReLU multilayer perceptron these transformations are exactly the permutations of hidden units and the positive rescalings of individual units: scaling a unit’s incoming weights up and its outgoing weights down by the same factor does not change what the network computes. Together they form the *unit symmetry group* of the architecture. By construction $T(B)$ computes the same function as B, that is $T(B)(x) = B(x)$ for every input x .

Proposition 1 (endpoint invariance). Define the *chord* as the straight line connecting the two endpoint loss values, $(1-\alpha)\cdot L(A) + \alpha\cdot L(B)$. It depends only on the endpoints and is the baseline used in the definition of the interpolation barrier; it is not the loss along the interpolation path in weight space. For every function-preserving T , the pair $(A, T(B))$ has the same endpoint losses as the pair (A, B) , and therefore the same chord. The interpolation path itself is generally not invariant: the losses along $(1-\alpha)\cdot A + \alpha\cdot T(B)$ change with T . This is exactly the room an alignment has to lower a barrier. The proof is immediate from the definition of function-preserving.

Scope of the alignment guarantee. The aligner used here is guaranteed to recover a permuted-and-rescaled copy of a network exactly. That is an important special case, but it does not prove that the alignment is optimal over the whole symmetry group for independently trained networks. Consequently the share of the barrier attributed to removable coordinate mismatch is a lower bound, and the residual share an upper bound, on their true values.

Proposition 2 (no merged model can serve both parents). Let h be any single classifier; in particular, any interpolated or merged model, under any alignment. On every input $x \in S$ the two parents disagree, $f_A(x) \neq f_B(x)$, so h must disagree with at least one of them. Writing $\varepsilon_P(h)$ for h ’s error rate against parent P ’s labels,

$$\varepsilon_A(h) + \varepsilon_B(h) \geq \mu(S), \text{ hence } \max(\varepsilon_A(h), \varepsilon_B(h)) \geq \mu(S)/2.$$

When two models’ conventions conflict on a set of mass $\mu(S)$, any hybrid of the two is wrong on at least one parent’s task at least $\mu(S)/2$ of the time. This floor is information-theoretic, holding regardless of the alignment group, the architecture, or the merging operator. In the fitness sense it is reproductive isolation: beyond a given functional conflict, no recombination operator can produce an offspring faithful to both lineages.

What remains empirical, and how the experiment is designed. Propositions 1 and 2 do not bound the single-task path barrier: the loss along the interpolation between A and $T(B)$, evaluated on one parent’s task alone. In principle such a path could dip toward one parent’s function and yield a low barrier even under conflict. Whether it does is an empirical question, and it is precisely what the experiment measures. The measured answer is that it does not. In the conflict condition the barrier is unchanged by permutation alignment (the **residual** readout) and by alignment modulo the full permutation-and-positive-rescaling group (the **residual_scale** readout), while the very same aligner removes almost all of the barrier between independently initialised networks, the positive control. Work on richer symmetry groups for transformers (83) strengthens the removable side of the decomposition and is therefore complementary to this result: the more barrier a larger group can remove for *compatible* models, the sharper the meaning of the barrier that survives for *incompatible* ones. Proposition 2 caps what any of these methods could ever achieve on the conflict set.

Terminology used in the paper. “Residual (after alignment)” denotes the estimated functional incompatibility: the part of the merge barrier that remains after the architecture’s unit symmetries have been divided out. For ReLU MLPs I align modulo the full unit symmetry group, so the estimate is not confounded by symmetries of that architecture class that the aligner might have missed.

SI Text S2. Emergent versus imposed incompatibility

The conflict condition *imposes* contradiction: the two label maps disagree on S by construction, which pins $\mu(S) > 0$ and activates Proposition 2. A genuine Bateson–Dobzhansky–Muller incompatibility is instead *emergent*. Each lineage’s substitutions are harmless on their own background, so the training signals never contradict and $\mu(S) = 0$; any incompatibility appears only when the two lineages are combined.

Two conditions realise this emergent setting. In **disjoint**, the parents are specialists on complementary classes. In **augment**, they learn divergent input conventions on the same task. Neither condition contains label conflict, so any barrier that survives alignment cannot be attributed to label conflict. Such a barrier would be the emergent-speciation signal proper.

Both readings were registered before the run. If the residual barrier grows with divergence, then model speciation is emergent in real weights, and the trajectory seen in the analytic speciation model is realised. If the residual stays at the level of the **shared** control, then within this regime trained networks are more merge-compatible than the biological analogy predicts. The second reading would bound the analogy, and be a useful design result in its own right: merging is safe whenever there is no functional conflict.

Outcome. Four replicates, with divergence up to 3,200 steps — up to $6.4\times$ the shared base training — returned the second reading. The residual barrier was 0.000 at every divergence in both emergent conditions. Merging moreover *rescued* the **disjoint** specialists, which had forgotten the classes outside their specialty: at the longest divergence the parents score 0.535 and 0.474 on the full task, while the merged model holds approximately 0.955 at every divergence tested. This is a sustained Fisher–Muller rescue at zero barrier. Within this regime, reproductive isolation in real weights required functional conflict. The same question at language-model scale is answered by the duration arm of the language-model speciation experiment, which likewise found no isolation from over-training alone (1 to 12 epochs); whether still longer horizons erode mergeability (cf. 83) remains open.

SI Text S3. Compatible loci and conflicting alleles in a multigenerational population

The two kinds of new knowledge. A *locus* is a position in the genome, and *alleles* are the alternative versions that can occupy it: one blood-group locus, three alleles A, B and O, of which any one chromosome carries exactly one. In a model population a locus is a slot for a capability (“how to answer a two-way question”) and alleles are the incompatible conventions that could fill it (“yes/no”, “true/false”, “1/2”). A skill that conflicts with nothing a lineage already holds occupies a new locus and is simply added; a skill that demands a different convention for a question shape the lineage already answers is a competing allele, and a single model, like a single chromosome, carries one. Proposition S2 gives the cost: when two parents’ conventions disagree on a share $\mu(S)$ of inputs, any merged child errs against at least one parent on at least $\mu(S)/2$ of them. In the six-generation population a lineage obliged to merge at generation $\mathbf{t}$ pays that floor against its partner’s conflicting conventions; because the child continues the lineage, the loss is inherited, and the next generation’s conflict adds to it. Under the Latin-square curriculum $\mu_{\mathbf{t}}(S)$ is zero while partners are complementary (their skills occupy disjoint loci) and becomes positive once a partner carries a differently conventioned version of a skill the lineage already holds. Two of the six families — yes/no questions and two-way pronoun resolution — have the most idiosyncratic conventions and were measured in calibration at 0.00–0.04 accuracy on every other family, so they carry the largest $\mu(S)$ against every partner; the generation at which the curriculum hands them to a lineage’s partner fixes when that lineage’s collapse begins.

Negative controls that isolate convention conflict. Three alternative explanations of the obligate arm’s collapse were tested directly and refuted. (i) *A destructive skill spreading through merges.* A single 50/50 merge of two clean single-skill adapters (science questions 0.838 / yes-no 0.000; yes-no 0.800 / science 0.300) scored 0.863 and 0.787, mean 0.825 against 0.550 for the better parent: one merge is protective, not destructive. (ii) *Geometric dilution of an adapter’s signal under repeated averaging.* Five chained convex merges left the first skill’s

accuracy unchanged even though its nominal weight fell to $1/32$; but a scaling control showed the adapter alone delivers nothing at $1/32$ (0.000; full effect down to $1/8$), so what propagated through the chain was the answer format supplied by whichever partner carried enough weight, not the skill. Dilution is refuted, and the transmitted quantity is identified as the convention. (iii) *Continued training on merged weights.* Merging then training on the incoming family beat merging alone on the tracked skill in four of five rounds and on the incoming skill in all five, and absorbed the one format shock that dropped the merge-only chain ($0.567 \rightarrow 0.883$). With capacity ruled out by the lifelong-editing benchmark (80) at three orders of magnitude more content, convention conflict is the mechanism that remains — the one the framework predicts, and the one single-model studies report (81, 82).

Neutral and functional variation. Three adapters trained on the same family, differing only in seed and data draw, were near-orthogonal in weight space (pairwise cosine $+0.006$) and disagreed on 24% of answers, yet merging two of them gave 0.887 against 0.800 for the better one — exactly the fraction of questions on which either was right (0.887). Decomposing the weight change across seeds, roughly 85% of a LoRA delta is run-specific: shared signal power 6.2 (after correcting the finite-sample mean for its own noise) against noise power 35. That is why raw weight distance predicted nothing in the main text’s controlled test: most of what it measures is the counterpart of *synonymous substitution* — sequence change without functional change — which averages out when adapters for the same skill are combined, while the fraction that conflicts lives in the answer conventions. Averaging same-skill adapters before crossing them with a different skill improved the cross modestly ($0.825 \rightarrow 0.850$) while leaving each single skill unchanged, the inbred-line pattern: averaging within a line does not improve the line, it makes it cleaner to cross.

Attenuation and the effectiveness cliff. Scaling an adapter’s weights down does not degrade its skill gracefully. Each skill holds full accuracy to a skill-specific fraction ($1/8$ for science questions, $1/4$ for reading-comprehension spans, $1/2$ for commonsense completion, $1/4$ for yes/no) and then loses nearly everything within one further halving. Four of six adapters scored higher when attenuated (inference $0.40 \rightarrow 0.68$ at $1/4$; completion $0.75 \rightarrow 0.82$ at $1/2$; spans $0.72 \rightarrow 0.78$ at $1/4$; science $0.87 \rightarrow 0.92$ at $1/8$): they were over-trained at full strength — the effect reported for merging experts (84, 85) — and recoverable here by one scalar per adapter with no retraining (six separate specialists $0.678 \rightarrow 0.755$). Denoising across seeds does not move the cliff, so the limit is signal magnitude rather than signal-to-noise. Choosing per-skill merge weights from these solo curves failed (0.686 – 0.689 against 0.708 for uniform weights): in a six-way merge a skill’s effective strength is its weight relative to the others — six conventions competing for one output — so raising one starves the rest.

SI Text S4. Proof of the blending-inheritance proposition

Setting. K parents; each independently retains a given rare item with probability q , and a parent that retains it assigns it mass p . The child draws n samples from a *source distribution* and keeps the item if at least one draw is that item. Two sources are compared: (A) one parent chosen uniformly at random; (B) the mean of the K parents’ distributions.

Expected mass is conserved. Let $J \sim \text{Binomial}(K, q)$ be the number of parents retaining the item. Under (A) the source mass of the item is p with probability q and 0 otherwise, so its expectation is pq . Under (B) the source mass is pJ/K , whose expectation is $p \cdot E[J]/K = pq$. The expected number of copies in the child’s sample, n times the source mass, is therefore npq under both schemes (linearity of expectation).

Survival agrees to first order. Write $f(x) = 1 - (1 - x)^n$ for the probability that at least one of n draws hits an item of source mass x ; f is increasing and concave, with $f(x)$

$= nx + O((nx)^2)$. Survival is $E[f(M)]$ with M the (random) source mass. Under (A), $E[f(M)] = q \cdot f(p)$; under (B), $E[f(M)] = E[f(pJ/K)]$. When $n \cdot p \ll 1$, every realised mass satisfies $nM \leq np \ll 1$, so $f(M) \approx nM$ and both expectations reduce to $n \cdot E[M] = npq$: the $1/K$ dilution of scheme (B) is cancelled exactly by the item being present in the mixture whenever any of the K parents holds it. (Equivalently, in this regime the child's copy count is approximately Poisson with mean nM , and Poisson thinning by $1/K$ composed with a K -fold union preserves the mean.)

Boundary 1 (common items). Away from the first-order regime the comparison is settled by Jensen's inequality. Both schemes give M the same mean pq ; scheme (A) puts all its variance in the two-point distribution $0, p$, and scheme (B) has strictly smaller variance for $K > 1$. Since f is concave, $E[f(M)]$ is larger for the less variable M , so averaging never lowers expected survival, and raises it once np is not small. The extinction probability $1 - f$ is convex, which is the form in which the main text states this boundary. The proposition is thus a statement about rare items, where survival is linear in mass; it does not claim averaging is harmful in general.

Boundary 2 (union operator). Let the child instead draw from the distribution that assigns each item the largest mass any parent gives it, renormalised. The item's source mass is then p whenever $J \geq 1$, an event of probability $1 - (1 - q)^K$, increasing in K for every $q \in (0, 1)$. Expected survival $(1 - (1 - q)^K) \cdot f(p)$ therefore rises with K in every regime, without a first-order restriction. The operator needs an oracle (a verifier) to say which parent holds each item most strongly, which is what routing supplies in the language-model tier.

Both statements are confirmed by simulation in Fig. S8, where mean-mixture survival is flat in K and the item-wise maximum rises with it.

SI Table S1: the claims ledger (status / assumptions / evidence / limits)

Claim	Status	Key assumptions	Evidence	Known limits
Population collapse in the inheritance model is Wright-Fisher drift	Closed form; the diagnosis itself is due to prior work	Knowledge is a categorical distribution; refitting means re-sampling	Closed forms reproduced to <0.5%	Real learners add a signed, architecture-specific estimator bias (measured)
Grounding behaves like immigration, and the critical real-data fraction is far below one	Closed form, plus the sign confirmed empirically	Fresh samples from a fixed, non-drifting truth	Exact H_{eq} ; $g^* \approx 0.048$; sign holds in RNN/MLP/VAE and on MNIST	Deepest tail unrescuable at feasible budgets ($m \sim 1/p$); sharp threshold softens in trained nets
“Merge, don’t average” conservation	Exact for the output-mean operator	Rare-item regime; an oracle/verifier identifies the strongest source	E4 closed form + simulation; neural reproduction	Weight-averaging and routing are empirical cousins, not instances; budgets differ; bridge = the headroom rule
Offspring exceed every parent (Fisher-Muller)	Interpretation + empirical	Complementary (decorrelated) parents; verifiable fitness	E8 (inheritance model); LoRA merges beat the best specialist overall in every seed at 0.5B (5 seeds) and 7B (3 seeds)	LLM tier: 3 lexically-distinct families
Outbreeding depression on rugged landscapes; operator design rule	Biological-model result; hypothesis at LLM scale	NK epistasis stands in for skill entanglement	E9–E10; directed selection rescues	Not yet mapped onto a real task-entanglement measure
Optimal mate-pool breadth shrinks with ruggedness	Biological-model result; hypothesis for merging populations	Ring population, local selection	E14	Phenomenon known to island-model evolutionary computation; the contribution here is the mapping and the diversity/mean decomposition
Merge failure decomposes into a coordinate artefact plus a functional residual	Empirical at the trained-network and language-model tiers	Alignment enumerates the architecture’s unit symmetries	Full-symmetry residual ≈ 0 for compatible parents versus $\approx$ the naive barrier under conflict; a cliff in hybrid fitness; function-specific breakdown at the LLM tier	Scoped to aligned linear interpolation; conflict floor is information-theoretic, not genetic
Epistasis (not divergence) sets the cliff; snowball onset	Biological-model result; hypothesis at the neural tier	BDM incompatibility structure	E12	Snowball count $\neq$ performance cliff without the effect-size link; neural test outstanding
Pre-merge functional disagreement predicts merge penalty	Empirical, within a controlled grid (0.5B, 13 conditions $\times$ 3 seeds)	Constructed conflict/overlap/duration axes; oracle-potential outcome (pre-registered; ordering sensitive to reference)	Clustered CIs exclude 0; held-out LOCO $\rho \approx 0.4$; selected geometry baselines ≈ 0	Head-to-head predictor differences not individually significant; only selected baselines; generalisation to real task pairs open
Confidence weighting improves rank prediction over raw disagreement	Not supported (pre-registered internal prediction)	—	Paired contrast over the same bootstrap resamples: $\Delta \setminus$	$\rho \setminus$
The predictor improves budget-matched operator choice	Open	—	Soup-vs-route gap readout noise-dominated at 0.5B	The practical payoff; untested
Emergent speciation without label con-	Not observed (pre-registered)	Shared ancestry; compatible tasks;	Residual 0.000 to $6.4 \times$ base training;	Bounds the hypothesis; longer hori-

SI Table S2: headline quantitative results

Headline quantitative results with sample sizes, uncertainty, and outcome definitions (full per-experiment tables and falsifier status in the per-experiment documentation).

Result	Setting / n	Outcome definition	Headline
Closed-form validation	Inheritance model; standing tests	Simulated vs closed-form H-decay, immigration equilibrium, multi-parent union	Agreement < 0.5%
Grounding retention	Inheritance model (E2); 100 lineages per grounding level	Fraction of equilibrium diversity retained at grounding g (operational threshold)	$g \approx 0.05$ retains $\geq 95\%$ in the tested setting; smooth in g
MNIST collapse & rescue	Conv-VAE, 4 replicates; frozen oracle (98.5% mode acc.)	Mode support / forward-KL over generations	Dry: 30→1 modes; 10% grounding: 30/30 held
Fisher–Muller in LLMs	5 seeds (0.5B) and 3 seeds (7B), fixed tests	Merged vs best-specialist accuracy (overall; worst family); $\pm$: 95% CI over seeds	0.5B ties 0.647 $\pm$ 0.027 vs 0.592 $\pm$ 0.009; 7B soup 0.873 $\pm$ 0.004 vs 0.807 $\pm$ 0.038 (soup – best +0.066 $\pm$ 0.036, 3/3 seeds)
Union vs blend (headroom)	3 seeds (0.5B hard); 3 seeds (7B hard)	Paired per-seed ordering, routing vs weight-average	0.5B: routing > blend in 3/3 seeds, one catastrophic blend failure avoided. 7B: routing 0.503 $\pm$ 0.007 vs soup 0.408 $\pm$ 0.021 (+0.094 $\pm$ 0.015, 3/3); soup vs best specialist +0.001 $\pm$ 0.041 (the seed-1 ‘soup below best parent’ did not replicate). Directed – soup +0.073 $\pm$ 0.031 (3/3)
Speciation decomposition	MLPs, 3 replicates	LMC error barrier residual after permutation+rescaling alignment	Same-task 0.001; conflict 0.497 (naive 0.502)
Emergent isolation	MLPs 4 reps to 6.4 $\times$ base training; LLM 1→12 epochs	Residual barrier; merged vs parent accuracy	0.000 everywhere; merge rescues parents (≈ 0.955 vs ≈ 0.50)
LLM speciation, seeds	0.5B; 3 training seeds; fixed test prompts	Conflict cliff: merge best-convention accuracy vs parents’ own at full conflict. Duration null: merged private-task accuracy, 1 → 12 epochs	Cliff in 3/3 seeds (merge 0.02/0.12/0.16 vs parents 0.23–0.25); merged coherence over the sweep 0.147 $\pm$ 0.013 → 0.100 $\pm$ 0.082. No isolation in 3/3 (0.760 $\pm$ 0.075 → 0.950 $\pm$ 0.010)
Predictive test	13 conditions $\times$ 3 seeds (0.5B)	Merge penalty vs oracle parent potential (pre-registered; $\pm$: clustered 95% CI)	Functional ρ +0.45/+0.46, CI excl. 0; LOCO $\rho \approx 0.4$; geometry n.s.; paired differences n.s.
Predictive test, seed sensitivity	Same; per-seed and leave-one-seed-out	Spearman ρ vs merge penalty within each seed alone (n = 13 conditions)	Functional +0.37 to +0.53 in every seed; weight geometry ≈ 0 in every seed; gradient alignment seed-unstable (–0.11 to –0.55)
Six-generation population	1.5B base; 3 lineages $\times$ 6 generations; 3 training seeds; fixed tests (60 per family)	Best-lineage accuracy over six families at the final generation (mean of seeds; per-seed contrasts)	Never merge 0.796; declinable merge 0.792 (Δ –0.03/+0.01/+0.01); forced stop after generation 2: 0.793 (declinable – stop –0.008/–0.006/+0.011); obligate merge 0.269 (declinable – obligate +0.57/+0.54/+0.45); own-ancestor merge 0.663; single model 0.802
Conflict-arrival curricula	Conflict-early / conflict-late (fixed tests)	Decline rate and obligate	Declines 0.56 → 0.78 (fixed tests)

SI Methods: experimental procedures

Every experiment in this paper is one YAML config, one runner invocation, and one artifact triple (`results.parquet` + the resolved config + a manifest carrying the master seed, git commit, library versions, and a content hash). The configs named below are the authority on any parameter; this section gives the scientific reasoning behind the choices. `REPRODUCING.md` maps each manuscript panel to the config and seed that produced it.

M1. Design principles

Four rules govern every choice that follows.

Test each claim at the cheapest tier that can falsify it. A closed form beats a simulation, a simulation beats a trained network, and a small network beats a language model, whenever the cheaper instrument can still return the answer “no”. A costlier tier is entered only where it adds a discriminating test rather than a replication — which is why several cells of the programme (Fig. 1A) are deliberately empty.

Match the precision of the claim to the precision of the instrument. The inheritance model is exact, so it carries the paper’s quantitative statements. Trained systems add optimisation noise and inductive bias, so at those tiers I claim signs and orderings, never magnitudes.

Make reality able to refuse. Every tier has an oracle that is independent of the model being measured: a fixed true distribution in the inheritance model, a lossless identity code or a frozen classifier in the neural tier, an exact-match verifier over procedurally generated tasks in the language-model tier.

Declare the falsifier before running. Each experiment states the outcome that would refute the claim it tests (per-experiment READMEs; SI Table S1). Two pre-registered predictions failed, and are reported as failures in the main text.

M2. Replication: what a replicate is, and how many

A replicate means something different at each tier, and conflating the three would misstate what the error bars cover.

In the inheritance model a replicate is an independent lineage: a fresh random stream driving the same resolved config, with sub-seeds derived from the master seed by `SeedSequence.spawn`. Because drift *is* the object of study, the spread across replicates is signal rather than nuisance, and replicate counts are set so that the confidence interval on the summary statistic is small relative to the effect being reported.

In the trained-network tier a replicate is an independent lineage including fresh weight initialisation and data ordering, so it carries optimisation noise on top of drift.

In the language-model tier a replicate is an independent *training* seed evaluated on *fixed* test sets. Holding the evaluation data constant while varying the training seed isolates training stochasticity, which is the quantity in doubt; it also means the seed-to-seed spread I report is not inflated by resampling the benchmark.

Replicate counts, and why each is what it is:

Experiment	Replicates	Reasoning
E1, E2, E3, E5, E6	100 lineages	Long horizons (400–600 generations) with drift-dominated variance; 100 lineages put the CI on stationary diversity well inside the effect being resolved
E4	200	Outcomes are per-item binary retentions, the highest-variance quantity in the paper
E7	20	Trajectory contrast (sexual vs asexual adaptation speed), large and monotone
E8	40	The vertical claim; the headline separation, so the most replicated of the genotype experiments
E9, E10	24	Landscape sweeps where each point aggregates 200 offspring internally
E11	12	Four-arm ablation over 80 generations; arms separate by margins far exceeding the CI
E12, E12_nk	15	Each point already averages 500 (E12) or 200 (E12_nk) offspring
E14	20	Breadth $\times$ ruggedness grid, 60 generations per cell
kernel_sharpen, kernel_smooth	24	Two-parameter kernel fits against neural reference endpoints
bridge	60	The harness gate: must detect <i>any</i> departure from the inheritance model, so the most replicated neural run
grounding	18	Nine-point grounding sweep with per-generation network retraining
collapse, architectures	5	Sign-level demonstrations across architectures; each lineage retrains a network 22–25 times
recombination	8	Operator contrast in trained weights
mnist_collapse	4	15 generations $\times$ a conv-VAE retrained from scratch each generation; the contrast (30 modes vs 1) is categorical
speciation_real, _cliff	3	Barrier decomposition; the quantity is a near-deterministic function of the training condition (residual 0.001 vs 0.497)
speciation_real_emergent	4	A null: replicates are spent on longer divergence horizons rather than more repeats
llm_merge_seeds	5 training seeds	The Fisher–Muller signature, the most-replicated language-model claim
llm_moe_hard_seeds, llm_directed_hard_seeds, llm_epistasis(+compat), llm_speciation_add	3 training seeds	Per-seed orderings reported individually rather than averaged
7B runs (llm_merge_hpc, llm_moe_hard_hpc, llm_directed_hard_hpc)	3 training seeds	Seeds 2–3 added 2026-09-11 (hpc/llm_7b_seeds.pbs, ~33 min per seed on one L40S); per-seed contrasts in figures/stats_llm_7b_seeds.py
llm_curriculum_v5_early,late(_obl)	3 training seeds each	Conflict-arrival curricula; per-seed contrasts and the pooled partial-correlation test
llm_curriculum_v5_cull	3 training seeds	Differential reproduction; per-seed

The asymmetry is deliberate: replicates are cheap exactly where the quantitative claims live, and the expensive tiers are asked only for the sign of an effect the cheap tier has already quantified. Where a single run is all there is, the manuscript says so.

M3. The inheritance-model tier

Knowledge is a distribution over K discrete items; reality is a fixed Zipf-tailed distribution $\mathbf{p}^*$; one generation resamples $\mathbf{n}$ draws from the parent, optionally mixes in $\mathbf{m}$ verified draws from $\mathbf{p}^*$, and refits. Implementation: NumPy/SciPy, no GPU, bitwise reproducible.

Parameter choices. $K = 500\text{--}1000$ with `zipf_s` = 1.1 and half the items designated tail: large enough that the rare tail contains hundreds of items (so tail statistics are not dominated by a handful of them) and small enough to sweep densely. $\mathbf{n} = 100\text{--}200$ sets drift strength; it is the population size in the Wright–Fisher correspondence and the distillation sample size in the AI reading. Horizons of 400–600 generations were chosen so that ungrounded lineages reach fixation and grounded ones reach stationarity within the run, which the trajectories confirm.

Sweeps. E2 sweeps grounding $\mathbf{g} \in 0, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.4$; E3 contrasts uniform against region-matched grounding allocation; E4 crosses parent count $K_T \in 1, 2, 3, 5$ with parent correlation $\rho \in 0, 0.25, 0.5, 0.75, 1$ and $\mathbf{g} \in 0, 0.02, 0.05$; E5 crosses selection mode (none / greedy / quality-diversity) with novelty weight; E6 compares four re-minting arms.

The correlated-parent construction (E4). Parent correlation is constructed directly rather than obtained by tuning drift, so that ρ is not confounded with $\mathbf{n}$, $\mathbf{m}$, tail size, or generation count. For each tail item a shared switch $\mathbf{z} \sim \text{Bern}(\rho)$, a shared retention $\mathbf{s} \sim \text{Bern}(\mathbf{q})$, and per-parent $\mathbf{u}^{(k)} \sim \text{Bern}(\mathbf{q})$ give parent $\mathbf{k}$ retention $\mathbf{s}$ if $\mathbf{z}$ else $\mathbf{u}^{(k)}$. This yields exact marginal retention $\mathbf{q}$ and exact pairwise correlation ρ , and is exchangeable, so ρ is a single scalar knob.

Multi-locus experiments (E7–E11, E14). Genotypes are $L = 12$ biallelic loci (4096 genotypes — effectively open-ended relative to the population sizes used), with fitness either additive or a Kauffman NK landscape whose interaction count K tunes ruggedness from 0 to 10. E9 and E10 breed from `n_parents` = 6 local optima into populations of 200 offspring; E10 additionally screens offspring and iterates (5 rounds, keeping 8). E11 runs a population of $N = 60$ agents for 80 generations at ruggedness $K = 8$, with mutation $\mu = 0.03$, 120 offspring per generation, and selection weighting true fitness against consensus conformity at $\mathbf{g} = 0.85$. E14 sweeps mate-pool breadth on a ring of $N = 48$ against ruggedness.

Speciation (E12). $L = 20$ loci, incompatibility density $\rho \in 0.1, 0.25, 0.5$, parental divergence swept 0–20 substitutions, 500 offspring per cell at recombination rate 0.5. E12_nk repeats the question on NK landscapes ($L = 16$, K 0–10, 40 parent pairs, 200 offspring).

Validation. Three closed forms are asserted as standing tests to within 0.5%: neutral heterozygosity decay $\mathbb{E}[\mathbf{H_t}] = \mathbf{H_0}(1 - 1/\mathbf{n})^t$, the exact immigration–drift equilibrium, and the multi-parent union formula. These run in CI alongside the correctness tests. If they fail, the science is wrong rather than merely the code.

M4. The trained-network tier

Why a synthetic universe. Measuring collapse requires knowing the true distribution exactly. Each mode is rendered as a token sequence carrying an identity segment (base-2 digits encoding the mode index losslessly, so the oracle reads the mode back with zero error) followed by style tokens drawn uniformly at random. The style segment gives genuine within-mode entropy, so a generative model must learn a distribution rather than memorise K fixed strings, while the identity segment keeps the measurement noise-free. Mode truth comes from the

same `make_true_distribution` used by the inheritance model, so “mode”, “region”, and “tail” denote the same objects at both tiers.

The bridge gate. Before any trained model is interpreted, a histogram generator is run through the identical harness; it must reproduce the inheritance model exactly. This separates harness bugs from model behaviour, and is why the bridge run carries 60 replicates.

Architectures and training. The recurrent generator is an embedding (26) $\rightarrow$ GRU (128 hidden; 192 in the architecture-generality run) $\rightarrow$ linear readout, trained each generation from scratch with Adam, learning rate 2×10^{-3} , batch size 256, 25 epochs, and evaluated by sampling 12,000–15,000 sequences. Feedforward and variational autoencoder generators share the harness. Retraining from scratch each generation (rather than fine-tuning) makes the generational step a clean refit, matching the inheritance model’s operator.

MNIST tier. Dataset: MNIST via torchvision (60,000 training images). Modes are digit class $\times$ stroke-thickness bin ($10 \times 3 = 30$ modes) with a Zipf frequency profile, so roughly eighteen modes are rare. The generator is a convolutional variational autoencoder (latent 32, $\beta = 1$), retrained from scratch each generation with Adam, learning rate 10^{-3} , batch 256, 30 epochs, on 6,000 images drawn from the previous generation’s own samples, for 15 generations, at $g \in 0, 0.1$. The oracle is a frozen two-convolution classifier trained once (5 epochs) combined with a deterministic thickness measure; it reaches 98.5% mode accuracy and its 30×30 confusion matrix is recorded in the manifest as the measurement floor. Build gates: the oracle’s accuracy, and generation-0 recovery of all 30 modes.

Speciation in trained weights. Two multilayer perceptrons (784–512–512–10, ReLU, no batch normalisation — batch statistics would break the permutation correspondence the analysis depends on) are forked from a shared base trained for 500 steps, then trained apart for 100–3,200 further steps (up to $6.4 \times$ the shared base) under SGD at learning rate 0.05, batch 128. Merges are weight averages; the readout is the linear-mode-connectivity error barrier before and after alignment. Alignment composes deterministic Re-Basin permutation matching with exact per-unit scale canonicalisation — the unit symmetry group of this architecture — and is gated by a control that must recover a permuted-and-rescaled copy exactly. Since the search space is that group rather than all possible alignments, the removable share is a lower bound and the residual an upper bound.

M5. The language-model tier

Base models. Qwen2.5-Instruct at 0.5B and 7B, open weights under a permissive licence, with the revision pinned. Using two sizes from one family makes scale the only variable that changes between the small and large runs; the 0.5B model carries five-seed protocols on the easy families; the 7B runs are replicated over three training seeds.

Task families, and why they are procedural. Three deliberately disjoint families — list operations, string transformations, and small-integer arithmetic — are generated procedurally from a seed. Procedural generation buys four things that a standard benchmark cannot: an exact-match verifier that plays the role of reality (an answer is right or it is not, with no judge model in the loop); freedom from train/test contamination, since every evaluation item is generated fresh from a disjoint seed offset; control over family disjointness, which is the precondition for specialists to be genuinely decorrelated parents; and a difficulty knob. A **hard** variant (multi-step list operations, Caesar ciphers and letter transforms, multi-step and larger arithmetic) exists because the easy families saturate a 7B base at ceiling, and saturation removes the headroom in which recombination operators can differ — a control that proved necessary, since two null results at 7B turned out to be saturation artefacts rather than scale effects.

Data splits. Training, validation, routing-calibration, and test items are drawn from non-

overlapping seed offsets by construction (test from 1000 + family index, routing from 2000 +, validation from 3000 +, training from the run seed). Test sets are fixed across seeds in the multi-seed protocols. Selection of merge weights uses validation only; the winners are then reported on the untouched test split.

Specialisation. Each parent is a LoRA adapter (rank 16, $\alpha = 32$) on the frozen base, applied to all attention and MLP projection matrices, trained with a manual supervised fine-tuning loop: answer-only cross-entropy (prompt tokens masked out of the loss), AdamW at 2×10^{-4} , batch size 8, 3 epochs, bfloat16, 400–800 training items per family. Low-rank adaptation is the right instrument here for a structural reason rather than a computational one: it confines each parent’s specialisation to an additive low-rank delta over an identical frozen base, which is what makes weight-space recombination between parents well defined.

Recombination operators. Fusion by uniform weight averaging (soup) and by sign-reconciled, magnitude-pruned task arithmetic (TIES); union by keeping specialists intact and selecting per input (oracle routing, and a training-free nearest-centroid router over the base model’s own prompt embeddings) or per module (winner-take-all by delta norm); and directed recombination, which breeds a population of Dirichlet-weighted merges, scores each on validation, and keeps the fittest.

Evaluation. Greedy decoding, exact match after canonicalisation. Alongside overall accuracy I report worst-family accuracy, because the Fisher–Muller claim is about competence across all families rather than an average that a single strong specialty can carry.

The controlled predictive test. Thirty-nine parent pairs (13 conditions $\times$ 3 seeds) span three axes that are decorrelated by construction: conflict (contradictory conventions on shared prompts, with private training budgets held fixed), compatible overlap (the same shared prompts under the same convention — overlap and volume without conflict), and duration (weight divergence with no conflict, 1 to 12 epochs). Six predictors are computed before any merge: confidence-weighted functional conflict, raw disagreement, gradient alignment at the shared base, LoRA-delta cosine and L2 distance, and a cross-task performance baseline. Probes are drawn blind to where the conflict lives. The outcome is the merge penalty against oracle parent potential, pre-registered, and also reported against best-parent and mean-parent references because the predictor ordering is sensitive to that choice.

The six-generation population. Base model Qwen2.5-1.5B (base weights, not the instruction-tuned variant; 0.006 accuracy on the families untrained). Six public datasets with per-family verifiers: natural-language inference (MNLI; label), science questions (ARC; letter), commonsense completion (HellaSwag; letter), reading-comprehension spans (SQuAD; normalised span with aliases), yes/no questions (BoolQ), and pronoun resolution (WinoGrande; 1/2). Each family’s pool is split into disjoint training, validation, and test items before any sampling, so validation and test never share an item. Two further curricula, conflict-early and conflict-late, are given as explicit orders: the two families whose answer conventions conflict (BoolQ yes/no, WinoGrande 1/2) occupy generations 1–2 or 5–6 of every lineage and the four compatible families fill the remaining generations in rotated orders, so adapter age and skill count rise one family per generation in both and only the arrival of conflict differs (`configs/llm/curriculum_v5_early,late.yaml`; obligate arms in the `_obl` configs; three training seeds each; `hpc/llm_curriculum_timing.pbs`). The conflict-timing readout is the partial Spearman correlation of the per-generation decline rate with an indicator of conflict presence, controlling for generation, with a seed-clustered percentile bootstrap (`figures/stats_llm_curriculum.py`). Differential reproduction (`cull: true`) applies truncation selection after each generation’s measurement: the lineage with the lowest all-families accuracy is re-founded from the one with the highest (adapter, taught families, example budget and ancestry archive are copied; the slot keeps its curriculum order; ties leave the population unchanged), recorded as `culled` and `cull_source` rows (`configs/llm/curriculum_v5_cull.yaml`; three training seeds; `hpc/llm_cull.pbs`). The second base lineage is HuggingFaceTB/SmolLM2-1.7B-Instruct

(Apache-2.0; Llama architecture), run through the unchanged `merge_seeds` and `moe_hard_seeds` protocols with its own adapter cache (`configs/llm/merge_seeds,moe_hard_seeds_smol.yaml`; `hpc/llm_smol.pbs`; `figures/stats_llm_smol.py`). Three lineages take the six families in a cyclic Latin square (each lineage’s order is the previous lineage’s shifted by two), which fixes partner complementarity — the fraction of the partner’s families a lineage has not yet seen — at 1.0, 1.0, 0.8, 0.67, 0.33, 0.0 across the six generations. Each generation a lineage draws 300 new items from its scheduled family and 150 replay items split evenly across families already seen; the child adapter (rank 16) is initialised from the parent’s and trained for 3 epochs at learning rate 10^{-4} (founders from the base at 2×10^{-4}). Recombination averages two adapters at each weight in 0.5/0.5, 0.3/0.7, 0.7/0.3; the winner is chosen on 20 validation items per family seen and then trains on the generation’s new family. In the declinable arm the unchanged parent is a fourth candidate scored identically. The contemporary partner is the next lineage in the square; the ancestor partner is the lineage’s own adapter three generations earlier; the self-replay variant draws its replay from the parent’s own answers rather than from the datasets. Reporting uses 60 test items per family. Because two of the families are binary, an accuracy threshold at 0.5 is chance, so the text reports mean accuracy over the families a lineage has been taught and the trajectory of its first-learned family rather than a count of families above a threshold.

The composed society. A population of N LoRA agents on a shared frozen base evolves for G non-overlapping generations. Each generation every agent answers a fixed validation pool (verifier scored) and a fresh conformity pool (whose modal answer defines the population consensus); selection scores agents by $\mathbf{g}\text{-fitness} + (1-\mathbf{g})\cdot\text{conformity}$; parents are chosen with or without a quality-diversity term over behavioural distance; offspring are bred by screened recombination; and each child is a fresh adapter distilled from its source model’s own answers, which makes the inheritance channel literally self-consuming. The verifier enters the loop only where $\mathbf{g} > 0$, but is used for reporting in every arm. The four-arm ablation removes grounded evaluation, recombination, or diversity preservation in turn.

M6. Negative controls

The design leans on controls that can remove a result rather than support one, and one of them did.

The compatible-overlap axis was added to the predictive test specifically to expose overlap-and-volume artefacts, and it did: the initial two-axis grid’s best predictor (LoRA-delta cosine, $\rho = +0.60$) collapsed to $\rho = +0.03$ once compatible overlap was present, identifying it as an artefact rather than a signal. The duration axis supplies divergence without conflict. The emergent-speciation condition supplies divergence with no conflicting signal anywhere, and returns a null. The budget-controlled speciation design (`conflict_mode: add`) removes the confound between conflict fraction and private training budget. The histogram bridge is a harness control. In the inheritance model, $\mathbf{m} = 0$ arms and $\rho = 1$ (fully correlated parents) are the null conditions against which the corresponding effects are read. In the six-generation population, three arms are controls — a single model taught the curriculum alone (no population), the never-merge population (no recombination), and merging with one’s own ancestor (shared conventions, partial complementarity) — and the self-replay variant of the obligate arm controls the replay channel; SI Text S3 records the three direct tests that refuted alternative mechanisms for the obligate arm’s collapse.

M7. Statistical procedures

Error bars on replicate means are normal-approximation 95% confidence intervals unless stated otherwise. For the predictive test, where rows share task-data seeds across conditions and are therefore not independent, inference uses a condition-clustered bootstrap (13 clusters, 4,000 resamples); predictors are compared by paired contrasts on the same resamples; generalisation is assessed by leave-one-condition-out prediction; and per-seed and leave-one-seed-out sensitivity are reported alongside, since three seeds cannot settle seed generalisation on their own. Outcome- reference sensitivity is reported rather than resolved. Where a difference is not significant at the sample size available, the manuscript says so rather than reporting the point estimate alone.

SI Statistics

Output of `figures/stats_llm_epistasis.py` (clustered CIs, paired predictor contrasts, LOCO held-out prediction, outcome-reference sensitivity, within/between-axis decomposition) — reproduced verbatim at submission. Chronology of the predictive test (prospective / adaptive / post-hoc) as disclosed in `results/llm_epistasis/README.md`.

SI Figures

Sixteen figures are cited from the main text by number. Each is the per-experiment figure regenerated from the committed results artifact (`figures/plot_*.py`), reproduced here without re-plotting, so panel titles still carry the experiment's working name. Five of them are inheritance-model results with no real-model counterpart in this paper, reported here because each reproduces an established result: blending versus union retention (Fig. S8), the Fisher–Muller super-parent (Fig. S9), outbreeding depression on rugged landscapes (Fig. S10), directed recombination (Fig. S11), and the mate-pool breadth optimum (Fig. S13).

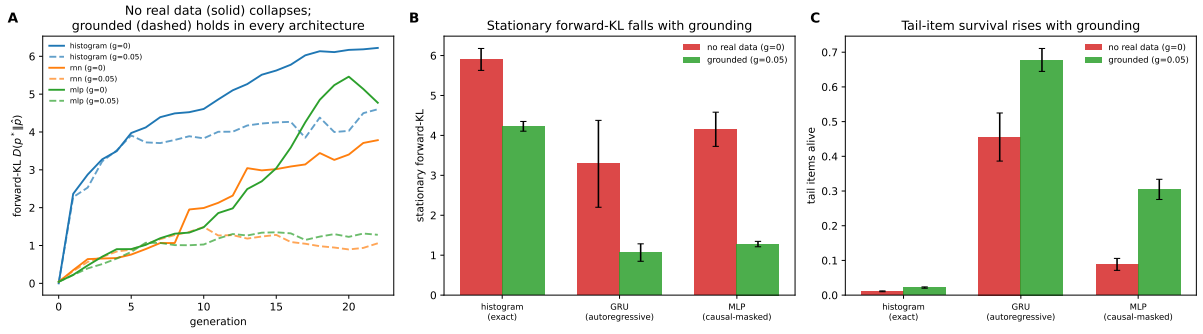


Figure S1: Collapse, and its arrest by real data, in three kinds of generator. The generational loop of Fig. 2 (train a child only on its parent’s output, with or without 5% real data) is run with an exact frequency count (a histogram, no network), a recurrent network and a feed-forward network, on a synthetic universe of 256 knowledge items whose true frequencies are known exactly; 200 samples per generation, 22 generations, 5 replicates. (A) Distance from the true distribution (forward KL divergence, which grows the more of the truth a model fails to cover) against generation: solid lines, with no real data, climb in every architecture; dashed lines, with 5% real data, stay low. (B) The same distance at the end of the run (error bars over replicates): real data lowers it in all three. (C) The fraction of rare items still alive at the end: real data raises it in all three. The histogram’s bars in C are small because a frequency count drops a rare item outright once it is unseen, whereas the networks keep some alive by smoothing (the subject of Fig. S2). A variational autoencoder was excluded because it failed the generation-0 fidelity check on this task.

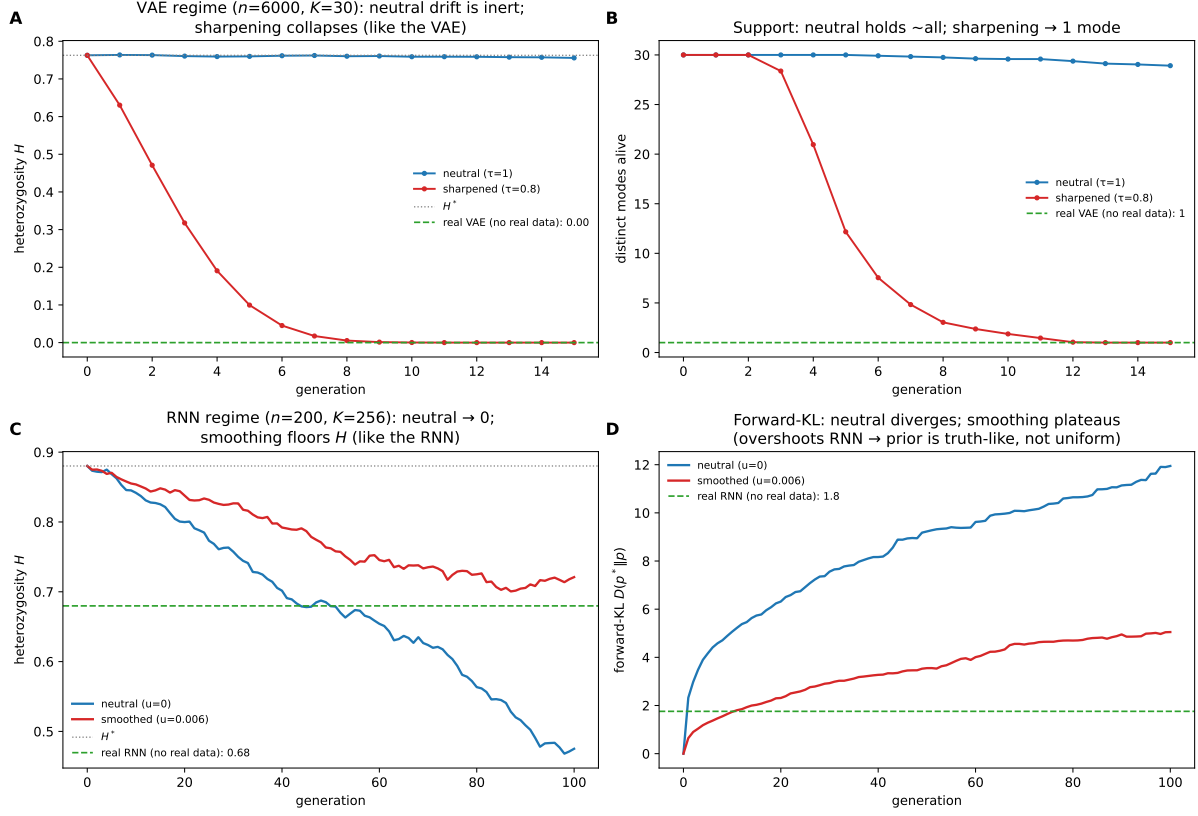


Figure S2: Why trained networks deviate from the ideal copier, in opposite directions. The inheritance model assumes a child's frequencies are exactly those it sampled from its parent. Two knobs are added to that copying step: a smoothing knob (a small pull toward treating every item as possible; mutation rate u) and a sharpening knob (a temperature $\tau < 1$ that concentrates probability on the commonest items). Blue: the ideal copier; red: the copier with one knob turned; green dashed: the level the real trained network reached with no real data; 24 replicates per regime. (A, B) The image network of Fig. 2 (6,000 samples per generation, 30 items): heterozygosity (A) and the number of distinct items still produced (B) against generation. The ideal copier barely drifts at this sample size, yet the real network collapsed to a single item; sharpening at $\tau = 0.8$ reproduces the collapse. (C, D) The recurrent network (200 samples, 256 items): heterozygosity (C) and forward KL divergence (D). The ideal copier drives diversity to zero, yet the real network keeps a floor near 0.68; smoothing at $u = 0.006$ reproduces the floor, though it overshoots the network's divergence (about 5 against 1.8), so the network's implicit prior is closer to the truth than a uniform one. In population-genetic terms the smoothing knob is recurrent mutation (variants appear without being inherited) and the sharpening knob is positive frequency-dependent selection (the majority gains, nothing new appears); a trained network behaves as drift plus one of these two biases, set by its architecture.

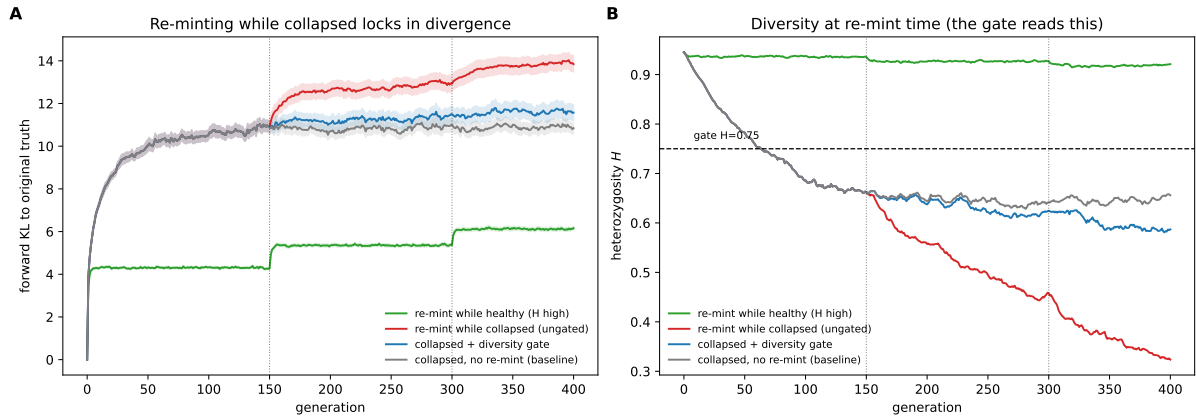


Figure S3: Re-baselining a collapsed population locks in its losses. A tempting shortcut is to declare a model's current output the new reference and discard the original data. In the inheritance model (500 items, 200 samples per generation, 400 generations, 100 replicates) the population's current frequencies are frozen as the new grounding reference at generations 150 and 300 (dotted verticals) and the original truth is kept only for measurement. Four arms: re-baseline while still diverse, under generous real data (green); re-baseline after collapse, under starved real data (red); the same starvation with re-baselining allowed only while heterozygosity is above 0.75 (blue); never re-baseline (grey). (A) Distance from the original truth against generation (bands over replicates): the red arm steps up at each re-baselining and never returns; the healthy arm shows small steps; the gated and never arms coincide. (B) Heterozygosity, with the gate's threshold dashed: the gated arm never re-baselines because it stays below the line. Once rare knowledge is gone from every copy it cannot be rebuilt (Muller's ratchet); a diversity gate prevents the shortcut from making the loss permanent.

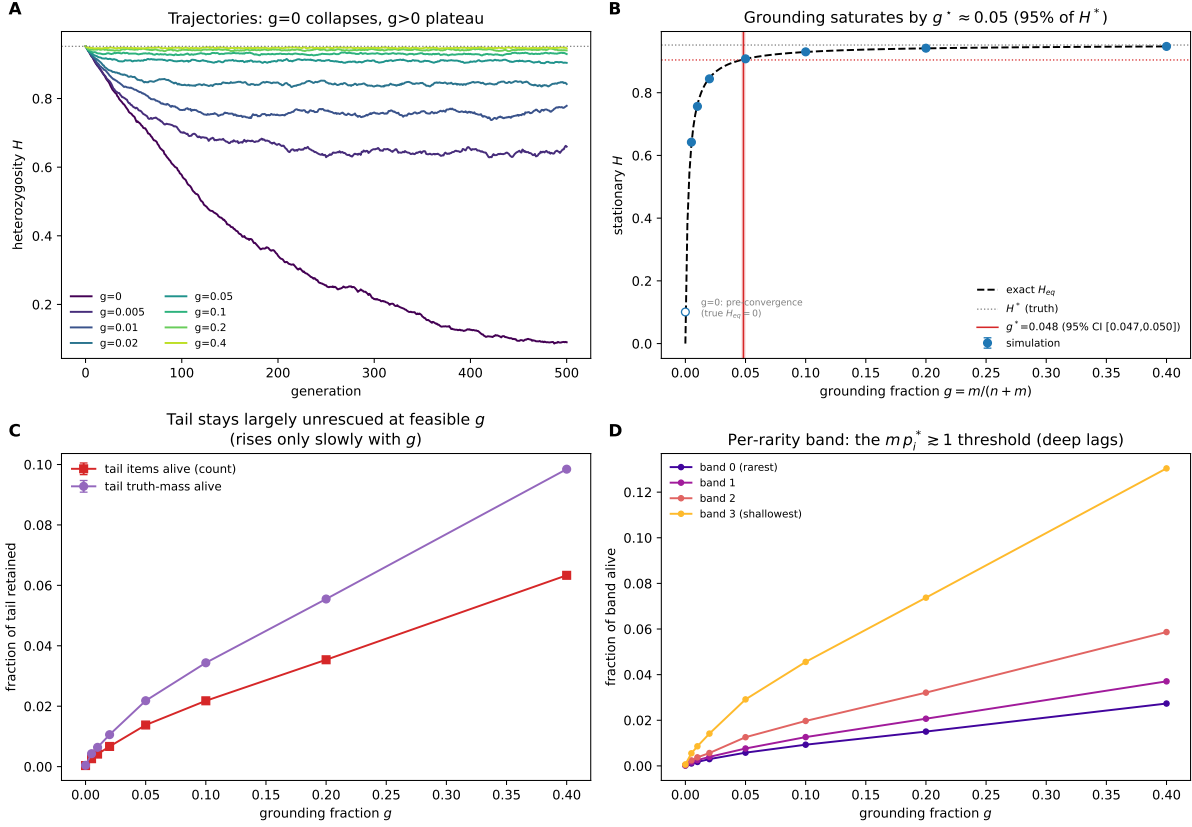


Figure S4: The full real-data sweep in the inheritance model (the experiment summarised in Fig. 2B). 1,000 knowledge items with a long tail of rare ones, 200 samples per generation, 500 generations, 100 lineages; each generation also receives m fresh real samples, so the real-data share is $g = m/(n+m)$, swept from 0 to 0.4. (A) Heterozygosity against generation, one line per g : with no real data it declines steadily; with any real data it levels off. (B) The level it settles at against g (points, simulation) with the exact prediction (dashed) and the real data's own diversity (dotted); the red line marks $g^* = 0.048$ (95% CI 0.047–0.050), where 95% of the real data's diversity is kept. The hollow point at $g = 0$ has not converged (its equilibrium is zero). (C) The fraction of the rare tail retained, counted by items (red) and by their share of the truth (purple): both rise with g but stay below 0.1 even at $g = 0.4$. (D) Survival by band of rarity, from the rarest (band 0) to the least rare (band 3): the rarest recover last. Overall diversity is cheap to protect; a rare item persists only once about one real example of it arrives per generation, so protecting it costs about one over its frequency in real samples.

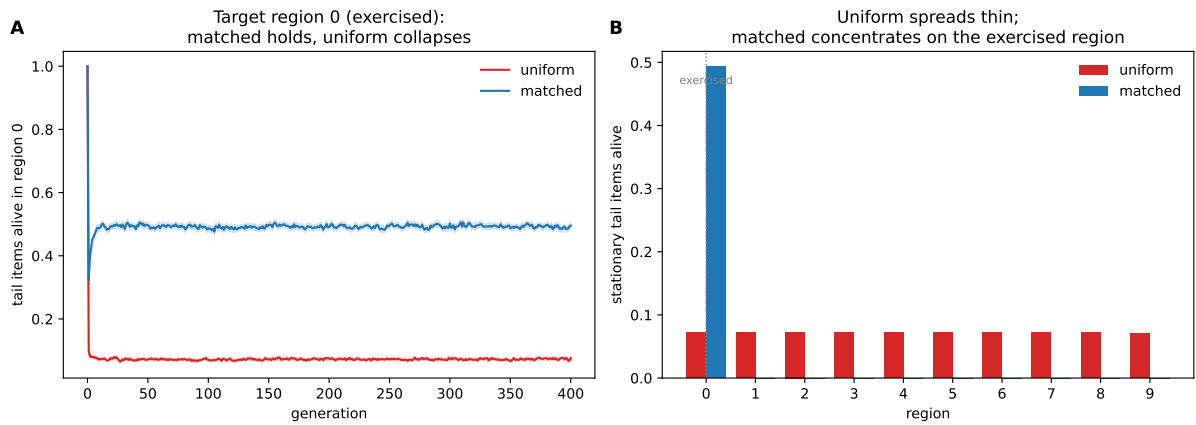


Figure S5: Real data protects only the topics it covers. The 1,000 items are divided into ten topics (regions) and the same total budget of real data is spent either evenly over all ten or concentrated on one topic the experimenter wants to protect; 200 samples per generation, 400 generations, 100 replicates. (A) The fraction of that topic's rare items still alive against generation, with real data aimed at it (blue) or spread evenly (red), bands 95% CI: aimed grounding holds about half the topic's rare items, spread grounding lets it fall to about 0.07. (B) Survival per topic at the end, same colours, the protected topic marked by the dotted line: aimed grounding protects its topic and leaves the others with no surviving rare items; spread grounding gives every topic the same low survival. Per-topic heterozygosity is confounded by how much of the truth each topic carries, so rare-item survival is the readout. A fixed budget of real data should be aimed at the knowledge one wants to keep.

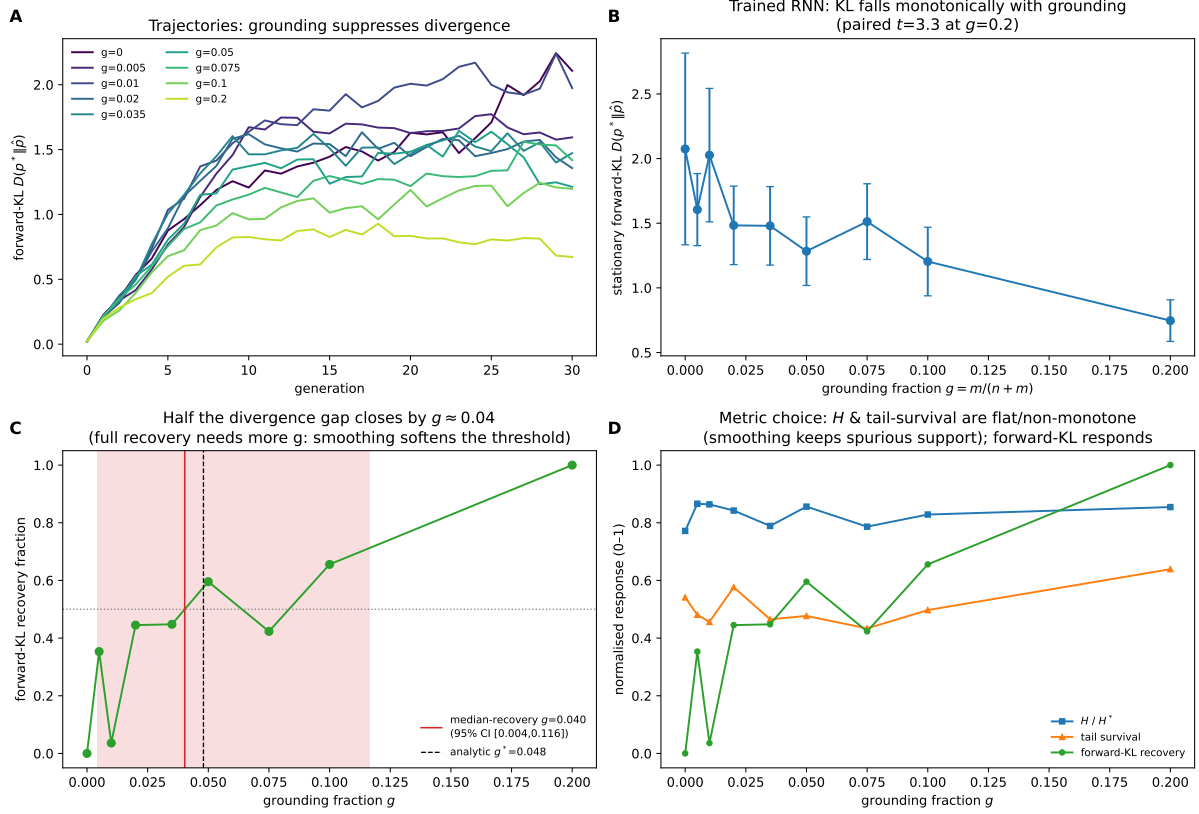


Figure S6: The real-data response in a trained recurrent network. The sweep of Fig. S4 repeated in a recurrent generator rather than the exact simulation: 256 items, 200 samples per generation, 30 generations, g swept over nine values from 0 to 0.2, 18 replicates. (A) Distance from the truth (forward KL divergence) against generation, one line per g : more real data suppresses the climb. (B) The final distance against g (error bars 95% CI), falling steadily from 2.08 with no real data to 0.75 at $g = 0.2$ (paired $t = 3.3$ at $g = 0.2$). (C) The fraction of the achievable improvement each g buys: half of it arrives by $g = 0.040$ (red line; bootstrap 95% CI 0.004–0.116 shaded), close to the simulation’s $g^* = 0.048$ (black dashed), but the full improvement needs g near 0.19. (D) Three ways of measuring collapse on one 0–1 scale: heterozygosity relative to the truth (blue) is flat near 0.8; the count of surviving rare items (orange) rises and falls with no pattern; the divergence-based recovery (green) rises cleanly. The direction of the effect matches the simulation, the threshold softens, and counting surviving items is the wrong ruler for a smoothing network, which keeps inventing rare items that are not in the truth; distance from the truth is the measure used for such networks.

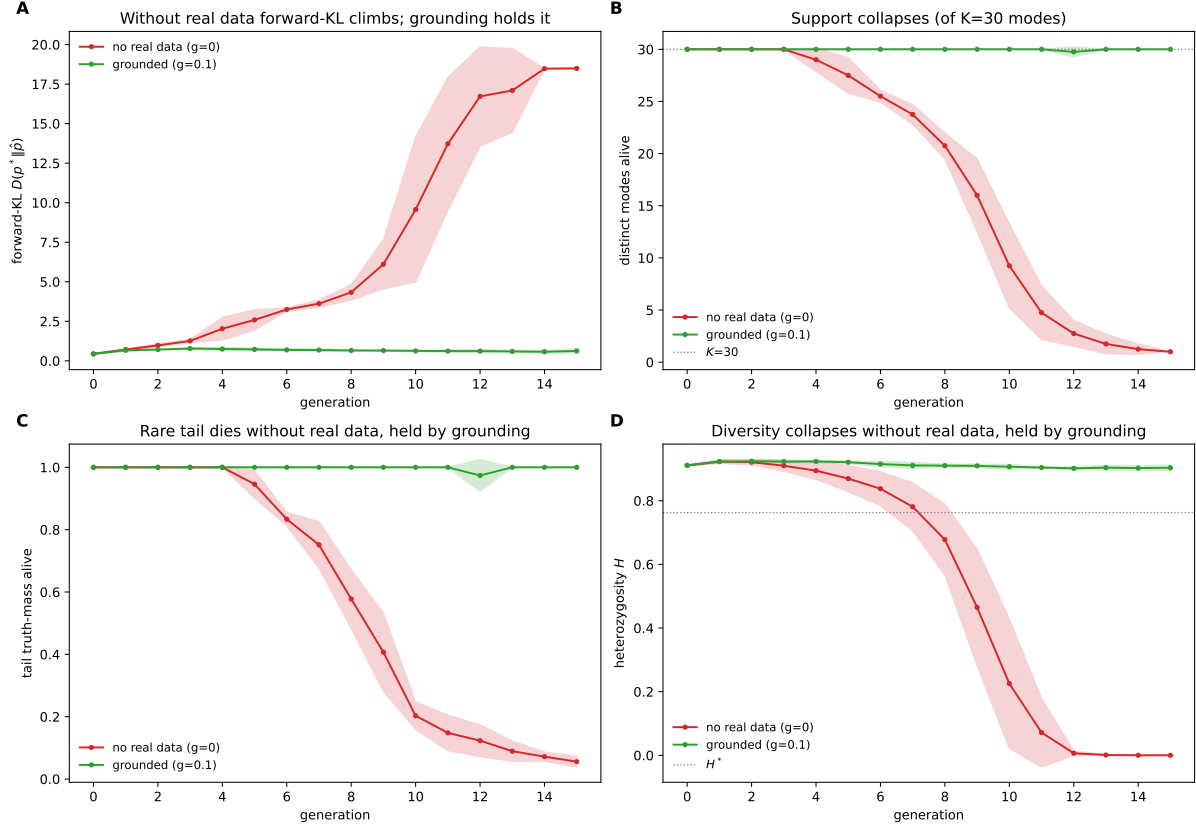


Figure S7: Collapse and rescue on real handwritten digits, in numbers (the experiment whose drawings are in Fig. 2A). A convolutional variational autoencoder is retrained from scratch each generation on the previous generation’s drawings plus a fraction g of real MNIST digits; the 30 kinds of digit (digit $\times$ stroke thickness, resampled to a long tail with about 18 rare kinds) are read out by a frozen classifier plus a thickness measure at 98.5% accuracy. Two arms, $g = 0$ (red) and $g = 0.1$ (green); 6,000 drawings per generation, 15 generations, 4 replicates, bands 95% CI. (A) Distance from the truth rises from about 0.5 to about 18 with no real data and stays near the floor with 10%. (B) The number of distinct kinds still drawn falls from 30 to about 1 with no real data; with 10% all 30 survive (dotted line). (C) The share of the rare kinds still alive falls to 0.06 with no real data; with 10% all of it is kept. (D) Heterozygosity falls to zero with no real data and stays near 0.9 with 10% (the truth’s value dotted). Everything the simulation predicts appears on real images with an independent judge; the dose of real data needed is about twice the simulation’s, for the reason shown in Fig. S2.

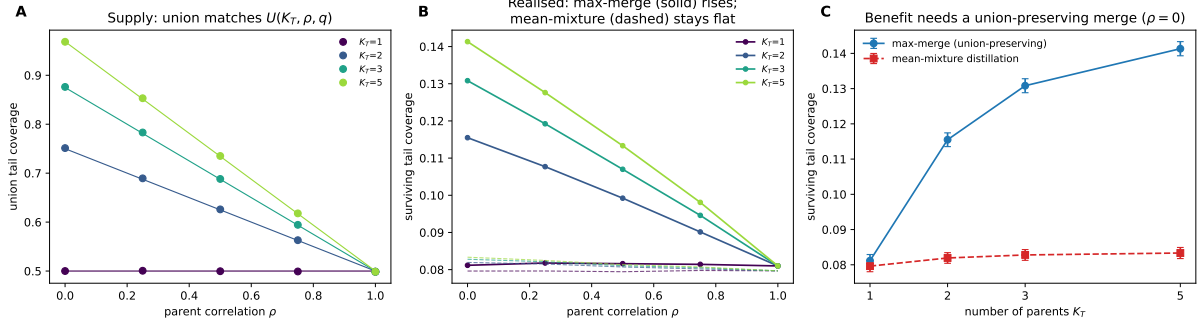


Figure S8: Averaging several parents cancels the benefit of having several; keeping each parent's strongest contribution does not. In the inheritance model (500 items) K_T parents each remember a random share of the rare items, with the similarity of their shares controlled directly by a correlation ρ (0 fully complementary, 1 identical); $K_T \in \{1, 2, 3, 5\}$, $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$, 200 replicates. A child is built either by averaging the parents' output frequencies or by keeping, for each item, the largest frequency any parent gives it (a union), and then resamples as every generation does. (A) The fraction of the rare tail held by at least one parent against ρ , one curve per K_T : points are simulation, lines an exact formula, and they match. (B) The fraction that survives in the child: solid lines (union) rise with more and less similar parents; dashed lines (averaging) stay flat near 0.08 whatever the number of parents. (C) The same at $\rho = 0$ against the number of parents (error bars 95% CI). Averaging dilutes each rare item by the number of parents, which exactly cancels the gain of having more parents to draw on (blending inheritance, the scheme Jenkin showed would swamp rare variants); the union realises the gain, and needs a judge to say which parent holds each item.

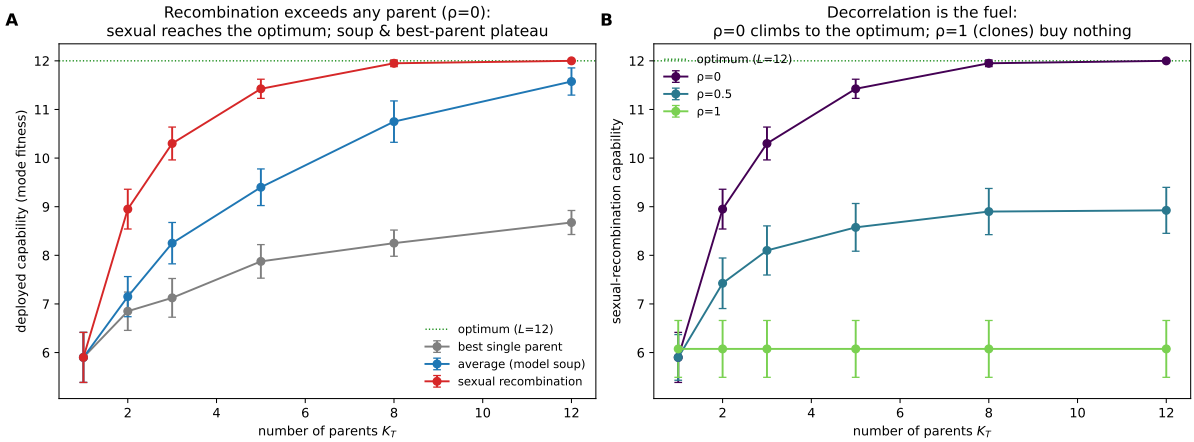


Figure S9: Many complementary parents can produce an offspring better than any of them. A capability is a string of twelve yes/no positions (a genotype of twelve loci) and fitness is the number of correct positions; each parent is a specialist, confident and correct (0.9) on the positions it has mastered and unsure (0.45) elsewhere, and no parent has mastered them all. Which positions a parent masters is drawn so that the number of parents K_T and their correlation ρ are independent knobs; the deployed capability is the fitness of the most probable genotype; 40 replicates, error bars 95% CI. (A) Capability against the number of parents when parents master different positions ($\rho = 0$): position-wise recombination (red) reaches the perfect score of 12 with eight parents; the best single parent (grey) sits near 8.7; the average of the parents (blue) reaches about 11.6 at twelve parents. (B) Recombination against the number of parents at $\rho \in \{0, 0.5, 1\}$: complementary parents climb to the optimum, identical parents stay flat near 6. This is the Fisher-Muller effect, unbounded because a model population is not limited to two parents; Fig. 3B is its counterpart in language models.

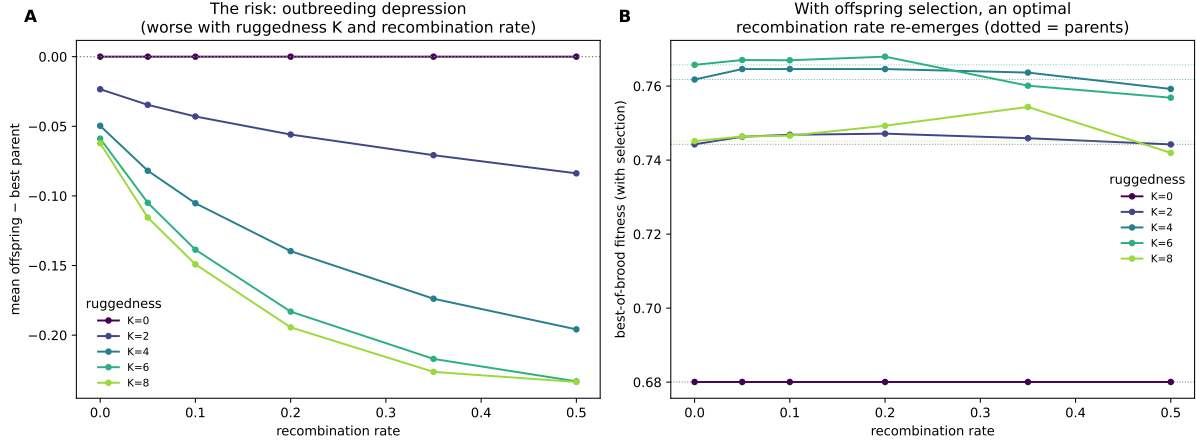


Figure S10: When skills are entangled, blind recombination harms the offspring. The twelve-position genotypes now sit on a rugged landscape (Kauffman’s NK model) in which a position’s value depends on its neighbours, with ruggedness K from 0 (positions independent) to 8 (highly entangled). Parents are local optima found by hill-climbing, the model of a trained specialist; offspring are made from them at recombination rates from 0 (copy a parent) to 0.5 (free shuffling); 24 replicate landscapes, 200 offspring per point. (A) Mean offspring fitness minus the best parent against recombination rate, one curve per K : on a smooth landscape the difference is zero; as K grows the curves fall, more steeply at higher rates, to about -0.23 at $K = 8$ under free recombination. (B) The fitness of the best offspring in a brood (parental level dotted): on rugged landscapes it peaks at an intermediate rate and falls back toward the parents under free shuffling. This is outbreeding depression; the optimal amount of recombination shrinks as skills become more entangled.

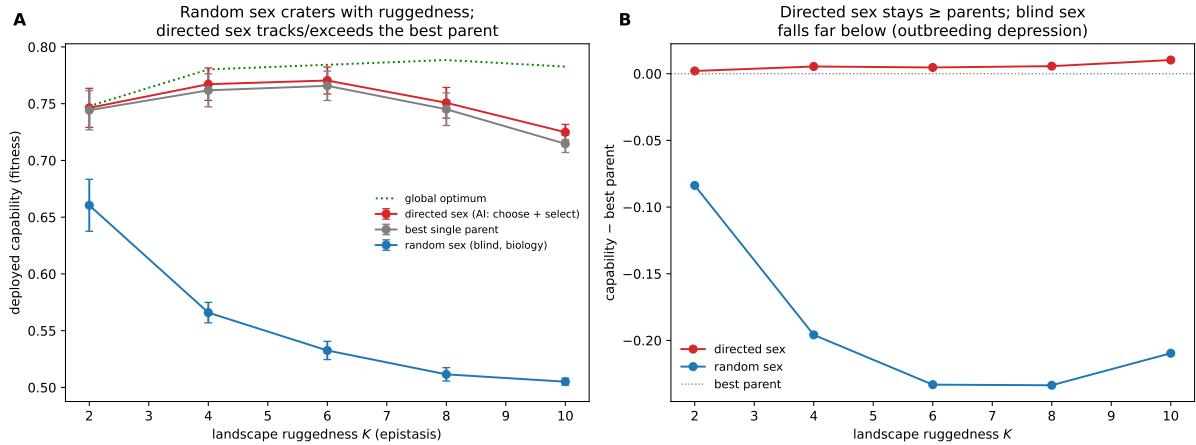


Figure S11: Choosing mates and screening offspring rescues recombination on rugged landscapes. On the landscapes of Fig. S10 three strategies are compared, all reported as deployed fitness in $[0, 1]$; 24 replicate landscapes, error bars 95% CI: the best single parent (grey); random recombination, as in biology (blue: random parents, free recombination, offspring taken as they come); and directed recombination, which a model population can do and a living one cannot (red: complementary parents chosen, many offspring generated at rate 0.2, the fittest kept, for five rounds). (A) Capability against ruggedness K with the global optimum dotted: random recombination falls from 0.66 at $K = 2$ to 0.51 at $K = 10$; directed recombination tracks the best parent and the optimum at every K . (B) The same as a difference from the best parent: directed stays at or above zero throughout; random falls to about -0.20 . In language models this is “breed many merges, keep the best” (Table S2).

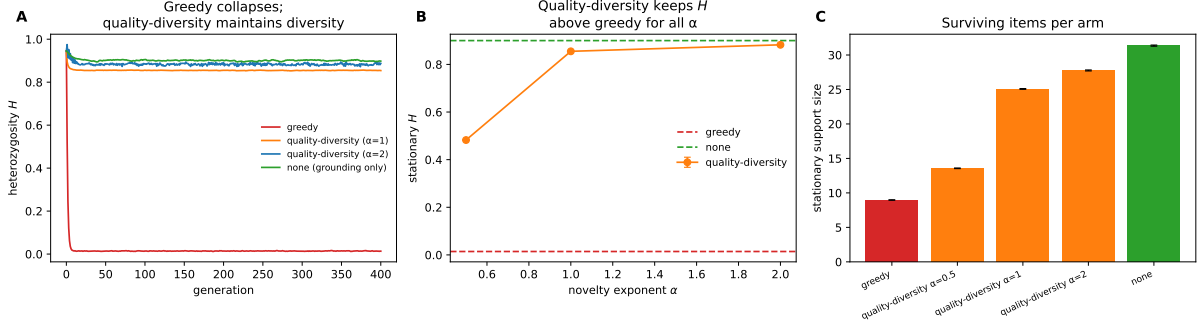


Figure S12: Selecting for the best destroys diversity; rewarding novelty preserves it. Each generation of the inheritance model (500 items, 200 samples per generation, 400 generations, 100 replicates, the same real data in every arm) now selects which items to keep, under three rules: no selection; greedy, keeping the items of highest true probability; and quality-diversity, which rewards an item for being rare as well as good, weighting item i by $f_i p_i^{-\alpha}$ with $\alpha \in \{0.5, 1, 2\}$. (A) Heterozygosity against generation: greedy (red) collapses within a few generations to about 0.01; quality-diversity at $\alpha = 1$ (orange) and $\alpha = 2$ (blue) and no selection (green) hold a plateau above 0.85. (B) The settled heterozygosity against α (orange), with greedy (red dashed) and no selection (green dashed) as references: it rises from about 0.48 at $\alpha = 0.5$ to about 0.88 at $\alpha = 2$. (C) The number of distinct items alive at the end: about 9 under greedy, 14 to 28 under quality-diversity, about 32 with no selection. Chasing the best outputs is a directional pressure on top of drift; diversity has to be an objective in its own right, which is the diversity-preservation ingredient of Fig. 4D–F.

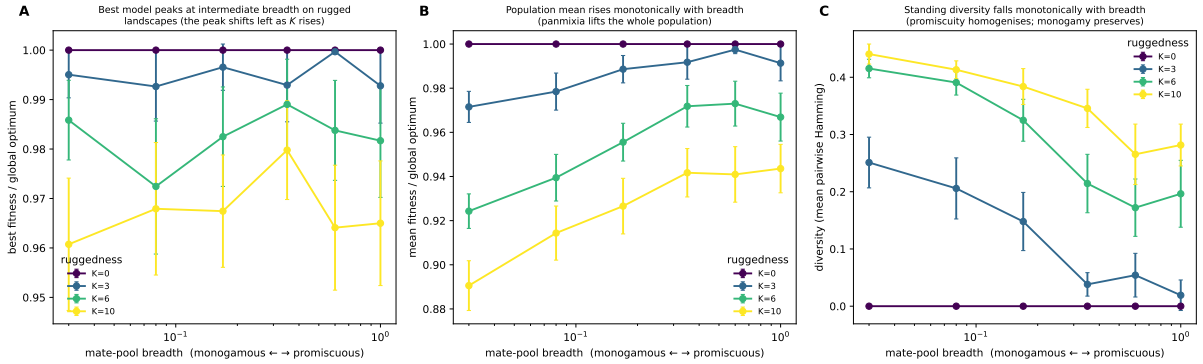


Figure S13: Who should mate with whom: the best mating breadth narrows as skills become more entangled. Forty-eight agents carrying twelve-locus genotypes sit on a ring and evolve for 60 generations on an NK landscape of ruggedness $K \in \{0, 3, 6, 10\}$; an offspring's second parent is drawn from a neighbourhood of half-width $\approx bN/2$, so the breadth b runs from mating only with neighbours ($b = 0.03$) to mating with anyone ($b = 1$), and an offspring replaces the agent at its position only if fitter (mutation 0.003, crossover rate 0.5, 20 replicates, error bars 95% CI, breadth on a logarithmic axis). (A) The best fitness reached, relative to the optimum, against breadth, per K : on a smooth landscape every breadth reaches the optimum; at $K = 3$ the best breadth is 0.6, at $K = 6$ and 10 it is 0.35, and mating with everyone falls below it. (B) The population's mean fitness rises with breadth at every $K > 0$. (C) Standing diversity (mean pairwise Hamming distance) falls with breadth, fastest on rugged landscapes. Wide mixing spreads a good variant fast but homogenises the population, so on entangled problems it loses the ability to explore several solutions in parallel (Wright's argument for structured populations).

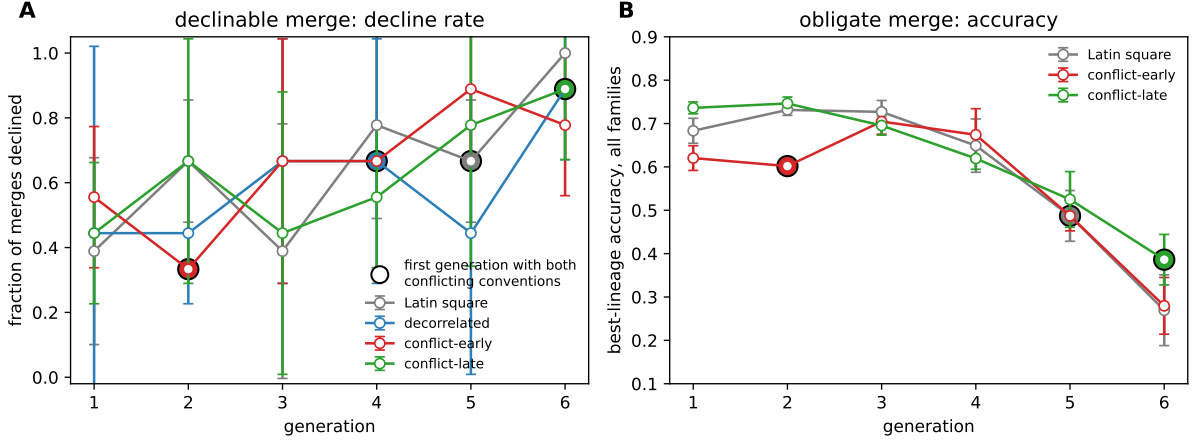


Figure S14: Conflict arrival does not set the timing of declines or of collapse. Four syllabi of the six-generation language-model population of Fig. 4 (three lineages, Qwen2.5-1.5B base, three training seeds each; mean $\pm$ 95% CI): the rotated syllabus, the syllabus with complementarity peaking mid-way, and two that differ only in when the two skills with clashing answer conventions (BoolQ yes/no, WinoGrande 1/2) arrive, in generations 1–2 (conflict-early) or 5–6 (conflict-late), the four compatible skills filling the rest in rotated orders so that adapter age and skill count rise one per generation in all four. The filled marker on each curve is the first generation at which both clashing skills are present in every lineage. (A) The fraction of proposed merges declined per generation in the declinable arm: declines rise with generation on the same schedule in every syllabus (partial Spearman with generation controlled: conflict present $\rho = -0.09$, 95% CI -0.45 to 0.15 ; generation $\rho = 0.45$; early and late pooled, $n = 36$). (B) Best-lineage accuracy over all six skills in the obligate-merge arm: the conflict-early population dips when the pair arrives, recovers by generation 3 and collapses from generation 5; the conflict-late population collapses from generation 4 with its clashing pair still to come. Final accuracy 0.28 (early) and 0.39 (late) against 0.80 and 0.78 for never merging. Moving the conflict by four generations moved neither the declines nor the collapse.

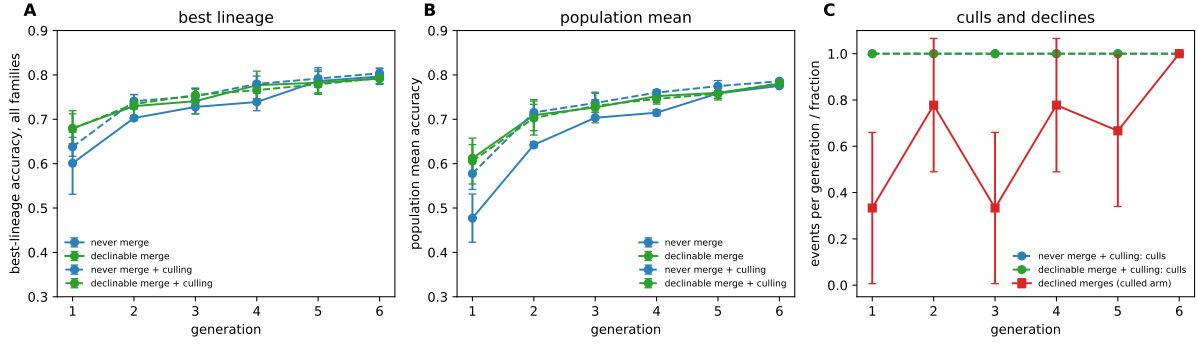


Figure S15: Survival of the fittest did not give merging lineages the edge. The rotated-syllabus population of Fig. 4 with truncation selection added: after every generation’s test the lowest-scoring lineage is re-founded from the highest-scoring one, keeping its own place in the syllabus ($N = 3$); the never-merge and declinable-merge arms are run with this selection (dashed) beside the same arms without it; three training seeds, mean $\pm$ 95% CI. (A) Best-lineage accuracy over all six skills: all four populations end within 0.01 of each other, near 0.80 (never merge + selection 0.804, declinable + selection 0.793, below in 3/3 seeds by 0.011 ± 0.003 ; 0.796 and 0.792 without selection). (B) Population mean over the three lineages: selection lifts the mean early (generation 2: 0.58 against 0.48 for the unselected never-merge arm) but the final means converge. (C) Selection acted every generation (one replacement per generation in every selected population, dashed) and declines in the selected declinable arm rose with generation as before. Recombination’s early lead (generation 1: 0.68 against 0.60) is present with and without selection and gone by generation 5 in both; under a syllabus that delivers every skill to every lineage, sex and selection each reach the same ceiling sooner and neither raises it.

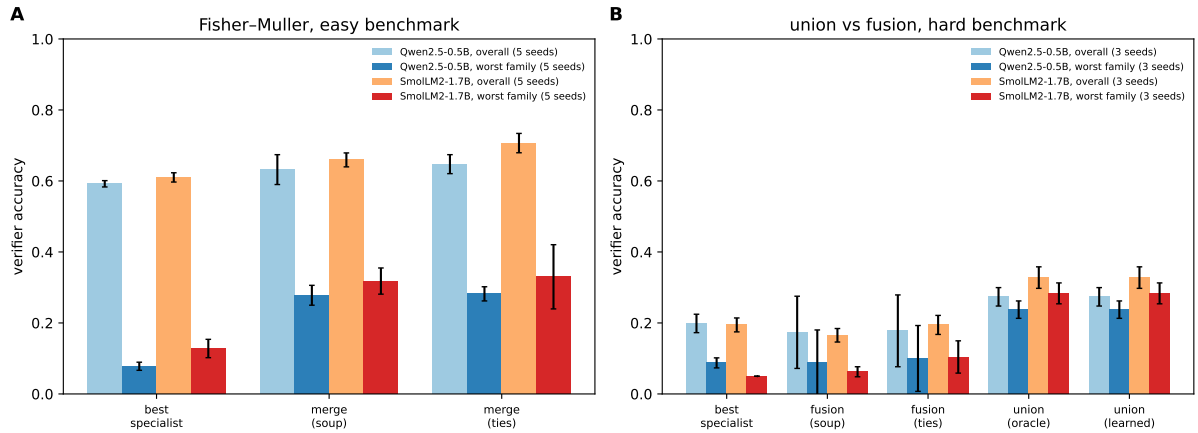


Figure S16: The two most-cited language-model results on a second, unrelated family of base models. The experiments of Fig. 3B and 3C re-run, with protocol, task families, test sets and seeds unchanged, on HuggingFaceTB/SmolLM2-1.7B-Instruct (Apache-2.0; Llama architecture; a different laboratory and pretraining corpus from Qwen), shown beside the Qwen2.5-0.5B-Instruct originals; bars are means over training seeds with 95% CI, overall accuracy (lighter) and worst-family accuracy (darker). (A) Easy tasks, five seeds per lineage: on SmolLM2 the averaged and interference-aware merges exceed the best single specialist in every seed (overall $+0.049 \pm 0.022$ and $+0.097 \pm 0.020$; worst family $+0.19$ and $+0.20$), matching the Qwen margins. (B) Hard tasks, three seeds per lineage: routing among intact specialists beats the weight average in every seed on both families, by a larger margin on SmolLM2 ($+0.162 \pm 0.036$ overall, $+0.221 \pm 0.029$ worst family), where the average falls below the best single specialist in every seed. The learned router equals the oracle router on both families because the families are lexically separable.