

Supporting 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

Generated from `paper/pnas/si.md`

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 (41) 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 be an honest bound on the analogy, and 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. 43) remains open.

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

Claim	Status	Key assumptions	Evidence	Known limits
Population collapse in the biological 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 (biological model); 7B LoRA merge beats every specialist on every family	LLM tier: 3 lexically-distinct families; replicated over five training seeds at 0.5B
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: $\Delta \backslash$	$\rho \backslash$
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 conflict	Not observed (pre-registered)	Shared ⁴ ancestry; compatible tasks; the divergences tested	Residual 0.000 to $6.4 \times$ base training; the merge rescues the specialists	Bounds the hypothesis; longer horizons/distribution shift/capacity pres-

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	Biological model; standing tests	Simulated vs closed-form H-decay, immigration equilibrium, multi-teacher union	Agreement $< 0.5\%$
Grounding retention	Biological model (E2); 100 lineages per grounding level	Fraction of equilibrium diversity retained at grounding $\mathbf{g}$ (operational threshold)	$\mathbf{g} \approx 0.05$ retains $\geq 95\%$ in the tested setting; smooth in $\mathbf{g}$
MNIST collapse & rescue	Conv-VAE, 4 replicates; frozen oracle (98.5% mode acc.)	Mode support / forward-KL over generations	Dry: 30 $\rightarrow$ 1 modes; 10% grounding: 30/30 held
Fisher–Muller in LLMs	5 seeds (0.5B), fixed tests; single 7B run	Merged vs best-specialist accuracy (overall; worst family)	Ties 0.647 ± 0.027 vs 0.592 ± 0.009 ; 7B 0.87 vs 0.77
Union vs blend (headroom)	3 seeds (0.5B hard); single 7B-hard run	Paired per-seed ordering, routing vs weight-average	Routing $>$ blend in 3/3 seeds; one catastrophic blend failure avoided
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 $\rightarrow$ 12 epochs	Residual barrier; merged vs parent accuracy	0.000 everywhere; merge rescues parents (≈ 0.955 vs ≈ 0.50)
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.

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 biological 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 biological 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 biological 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 biological 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_speciation	1	Single-run confirmations at a scale where each run costs GPU-hours; reported as sign-level and labelled as single runs

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 biological-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 teacher 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). Teacher 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-teacher $\mathbf{u}^{(k)} \sim \text{Bern}(\mathbf{q})$ give teacher $\mathbf{k}$ 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 $E[H_t] = H_0(1 - 1/\mathbf{n})^t$, 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 biological 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 biological 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 (24) $\rightarrow$ 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 biological 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 the multi-seed protocols and the 7B model the single-run confirmations.

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 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 $\mathbf{g} \cdot \mathbf{fitness} + (1 - \mathbf{g}) \cdot \mathbf{conformity}$; 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 re-

turns 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 biological model, $m = 0$ arms and $\rho = 1$ (fully correlated parents) are the null conditions against which the corresponding effects are read.

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

One per experiment, regenerated from committed artifacts: E1–E14, bridge/collapse/grounding/architectures/recombination, kernel (sharpen/smooth), mnist_collapse (+ montage), speciation_real (decomposition/cliff/emergent), llm_merge(_hpc/_seeds), llm_moe(_hpc/_hard_hpc/_hard_seeds), llm_directed(_hpc/_hard_hpc/_hard_seeds), llm_speciation(_add), llm_epistasis(_compat).