

# 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 · lab.gilest.ro

PNAS draft — generated from `paper/pnas/main.md`

## Significance statement

Artificial intelligence increasingly consists of populations of models rather than single systems. Models are fine-tuned from common ancestors, trained on data that earlier models generated, and combined by weight merging. These practices couple model generations the way reproduction couples biological generations, and they raise the same question: how does a population retain and accumulate abilities over time? I transfer the population genetics of sexual reproduction to this setting and test it in simulations, small neural networks, and language models. The framework recasts continual learning at the population scale and yields design rules: how much real data retraining requires, when to combine models, when to keep them separate, and how to anticipate a failed combination before making it.

## Abstract

AI development increasingly resembles a population process. Models are specialised, retrained on model output, and recombined by weight merging, and the practice is described in evolutionary vocabulary with little use of evolutionary theory. I treat multigenerational model populations as systems whose inheritance, diversity, and compatibility must be managed, and transfer the quantitative framework of the evolution of sex. Its starting point, that training on model output is genetic drift and model collapse its signature, is by now established from several independent directions; I develop the structure that follows from it. In a minimal inheritance model that is exactly Wright–Fisher, and measurably Wright–Fisher plus estimator bias in trained networks, I derive and test remedies. Grounding acts as immigration: a real-data fraction far below one retained most equilibrium diversity, with a per-capability observation floor that makes the rarest knowledge expensive under unstratified sampling. Refitting a child to the mean of its parents’ output distributions cancels the multi-parent gain to first order in the rare-item regime; union-preserving operators realise it. Merged language-model specialists exceeded every parent in replicated experiments. Blind recombination fails on rugged task landscapes; screening candidate offspring restores the gain. The optimal mating breadth narrows as skills entangle. Finally, I introduce model speciation: a merge barrier remaining after permutation-and-rescaling alignment tracks functional conflict, isolation did not emerge from compatible specialisation, and in a controlled test pre-merge functional disagreement predicted merge damage while weight-geometry baselines showed no detectable association.

---

# Introduction

Machine learning has become a population-scale phenomenon. Public repositories host millions of models (Hugging Face alone grew past three million by 2026), and these are not independent creations: the overwhelming majority are fine-tunes, distillations, or merges of a small number of foundation models, forming large family trees whose lineage structure, inherited traits, and mutation dynamics are already being mapped with explicitly phylogenetic methods (1–3). Weight-space *model merging*, the direct combination of trained parents into a new model, is mainstream community practice with standard tooling and thousands of hybrid checkpoints, including leaderboard-topping ones (4–7), and the engineering literature describes it in evolutionary vocabulary: “crossover,” “mutation,” “mate choice,” populations of merging models that climb benchmarks (5, 8–10).

The generations are coupled through data as well as through weights. Successive models increasingly learn from model output rather than from fresh human experience: frontier alignment pipelines are now predominantly synthetic (over 98% in documented cases; 11, 12), self-generated instruction data seeds whole lineages of descendants (13), a large and growing share of the public web is machine-generated or machine-translated text (14, 15), and the stock of human text is projected to be exhausted by frontier training within this decade (16). Meanwhile persistent multi-agent systems and emerging agent economies put many interacting models into sustained contact (17–20). A population whose members inherit from one another, recombine, and retransmit under these conditions is an evolving population in the technical sense, and that observation motivates this work. Here I transfer the quantitative framework of the branch of biology built for exactly this situation, the population genetics of the evolution of sex, and use it to treat multigenerational model populations as systems whose inheritance, diversity, and compatibility can be measured, predicted, and managed.

Training each generation of a model on the previous generation’s output degrades it (*model collapse*): rare capabilities vanish first, and the lineage drifts toward its own most common behaviour (21). That degradation is, mathematically, *genetic drift*, the loss of rare variants that any finite population suffers when each generation is a finite sample of the last. The identification has been made repeatedly and independently: for sequential inference chains before deep learning (22), for language-model text ecosystems (23), as a closed-form first-extinction law placing collapse onset at the Wright–Fisher first-extinction time (24), and in quantitative-genetic form for self-consuming diffusion models (25). A diagnosis reached so often, from such different starting points, marks population genetics as the natural mathematics of the setting, though only as its entry point: population genetics is not, at heart, a theory of decay; it is a theory of the mechanisms that maintain and build populations despite decay (immigration, recombination, selection, population structure) and of where those mechanisms reach their limits. This paper develops that fuller structure for model populations: the arc from drift through its remedies to its limit, reproductive isolation, carried as one framework from closed forms to trained networks to language models.

An operator of a model population faces recurring decisions for which there is no principled guidance: how much verified real data does retraining need before a lineage decays; will combining two particular models compose their abilities or damage them; can incompatibility be detected before paying for a failed merge; and when should specialists be kept separate rather than consolidated? In practice these are settled by convention and by trial-and-error search. They are also, recognisably, machine learning’s oldest problem at a new scale: *continual learning*, the struggle to acquire new abilities without losing old ones (26, 27), transposed from a single network to a population whose members inherit from one another. Population genetics, I will argue, prices these decisions. Table 1 summarises the correspondences on which the argument runs; the sections that follow develop them from closed-form theory to experiments in

trained networks and language models.

## The minimal model, and where its exactness ends

Knowledge is modelled as a distribution  $\mathbf{p}_t$  over  $K$  discrete items (capabilities, facts, modes of behaviour), with a fixed true distribution  $\mathbf{p}^*$  whose rare tail carries the knowledge most at risk. One generation is: \*draw  $\mathbf{n}$  samples from the parent’s distribution, optionally mix in  $\mathbf{m}$  verified real samples (“grounding”,  $\mathbf{g} = \mathbf{m}/(\mathbf{n}+\mathbf{m})$ ), and refit the child. *In this minimal inheritance model the resampling step is\** the Wright–Fisher process: the same equations, which I exploit as an engineering gate: the simulator reproduces the classical closed forms (heterozygosity decay  $E[H_t] = H_0(1 - 1/n)^t$ ; the exact immigration–drift equilibrium; the closed-form multi-teacher union) to within 0.5%, and these are standing tests in the codebase, not one-off checks.

The boundary of the exactness matters, and I measured it rather than assumed it. Real training adds approximation, optimisation noise, and inductive bias, and when trained networks are fit against the exact drift null they deviate in *opposite, architecture-specific* directions: a smoothing recurrent network resists collapse (keeping spurious variants alive), while a sharpening image generator accelerates it. A one-parameter *learning kernel* (a smoothing knob and a sharpening knob on the refit) reproduces both. Throughout, a real learner is therefore treated as Wright–Fisher *plus a signed, measurable estimator bias*, and the drift signs (rare-first loss; the grounding response) survived that bias in every architecture I tested, including a convolutional VAE retrained on its own generated digits, where the dry lineage collapses to a single blurred digit class while 10% grounding holds all thirty modes (Fig. 1). Retraining on a single parent is *asexual reproduction*, and sustained loss under it carries the defining consequence of *Muller’s ratchet* (28): once every copy of a rare capability is gone from all parents and sources, no recombination can rebuild it, so remedies must act before fixation-by-loss (a consequence-level correspondence; the minimal model lacks the ratchet’s recurrent-mutation driver).

**Table 1.** The dictionary. Each correspondence is stated with the level of support it currently has (exact = closed form in the minimal model; empirical = measured in trained systems; hypothesis = stated with a falsifier, untested or unconfirmed). The full claim-by-claim ledger with assumptions and known limits is SI Appendix, Table S1.

| Population genetics                            | Model populations                                       | Support                                                                               |
|------------------------------------------------|---------------------------------------------------------|---------------------------------------------------------------------------------------|
| Genetic drift in a finite population           | Training on finite samples of model output              | Exact (minimal model); signs in trained nets; diagnosis conceded to prior work        |
| Immigration from a fixed source                | Grounding with verified real data                       | Exact equilibrium; signs in RNN/MLP/VAE/MNIST                                         |
| Muller’s ratchet (asexual decay)               | Irreversible arm of model collapse                      | Correspondence, scoped: applies to unrecoverable loss                                 |
| Recombination / sexual reproduction            | Model merging                                           | Empirical at 0.5B–7B                                                                  |
| Fisher–Muller effect                           | Merged specialists exceed every parent                  | Analytic model; replicated in LLMs                                                    |
| Outbreeding depression under epistasis         | Merging entangled skills harms offspring                | Analytic model (NK landscapes); hypothesis at LLM scale                               |
| Mating systems / population structure          | Who merges with whom (breadth of the parent pool)       | Analytic model; hypothesis for real populations                                       |
| Reproductive isolation (BDM incompatibilities) | Merge failure from functional conflict                  | Empirical (MLP + LLM tiers, conflict-associated); emergent form not observed          |
| Selection on a fitness function                | Verifier-anchored selection (“reality that can say no”) | Analytic model (complementary with recombination and diversity in the tested society) |

## Results

### Grounding is immigration: cheap, with a floor

In the minimal model, grounding from a fixed real source is *immigration* into a drifting population (29–31), and the equilibrium diversity has a closed form the simulator matches exactly. That equilibrium is *smooth* in the grounding fraction (there is no phase transition in aggregate diversity), so the practical number is an operational threshold, and I define it as such: under the tested population size and Zipf source distribution,  $g \approx 0.05$  retained most ( $\geq 95\%$ ) of equilibrium diversity indefinitely, with the required fraction depending on sample size, source distribution, and the chosen retention target (dependencies in SI). Verified real data remains, on any of these definitions, cheap insurance at fractions far below one. But the same analysis yields a floor the field’s average-loss framing misses: under unstratified sampling from the source, a capability of rarity  $p$  appears in a real-data batch of size  $m$  with probability  $1 - e^{-m \cdot p}$ , so  $m \cdot p \approx 1$  marks roughly a 63% chance of one example per batch: a soft observation floor, with higher confidence priced accordingly, and with distinct consequences for continuous retention, stationary occupancy, and reintroduction after loss (immigration can restore an absent item; SI separates these). Protecting the rarest knowledge under unstratified grounding is therefore priced per item at cost  $\propto 1/p$ ; targeted or stratified sampling changes that cost, and recombination can recover rare capabilities *that are still retained across complementary parents* (next section). In trained networks the *sign* of the grounding response transfers everywhere I looked, with two deviations, both traced to the estimator bias above: sharp thresholds soften, and support-counting metrics decouple from truth (forward-KL is the operative collapse metric for a smoothing learner). On real images (Fig. 1B), dry self-training collapses a convolutional VAE to one mode while  $\sim 10\%$  grounding holds all thirty (the trained model needs roughly twice the exact-operator fraction, the measured price of the estimator bias).

## Recombination: a conservation law, its operators, and offspring that exceed every parent

The largest returns from the transfer concern merging. **Proposition (blending inheritance, rare-item regime).** Let  $K$  parents independently retain a rare item (mass  $p$  when retained), and let the child draw  $n$  samples either from one parent chosen at random or from the *mean of the parents' output distributions*. Expected item mass is identical under the two schemes; and in the rare-item regime  $n \cdot p / K \ll 1$ , where per-item survival is first-order in sampled mass, expected *survival* is also identical: the  $1/K$  dilution of averaging cancels the  $K$ -parent union gain to first order, so in this regime adding parents through the output-mean does not increase expected tail retention. Two boundaries: outside that regime, survival is a convex function of mixed mass, so the variance reduction from averaging can *reduce* extinction relative to a randomly chosen single parent; the cancellation is a first-order result about rare items, not a universal impossibility; and the contrasting union operator (keep each item's strongest source, then renormalise, which itself redistributes mass and presupposes a verifier or oracle to identify the strongest source) increases expected retention with  $K$  in all regimes in the minimal model. The practically important operators, *weight averaging* (a nonlinear network's weight-mean does not compute its parents' output-mean) and *routing among intact specialists* (32) (different storage and inference budgets from a single child), are its empirical cousins, and the measured bridge is a *headroom rule*, stated qualitatively: in language models, union-preserving operators beat the weight-average where that average falls short of attainable performance, and add nothing where it does not (easy-versus-hard contrasts at two scales; a quantitative form of the relationship is untested). On easy tasks a capable base's average is already at ceiling and refinements add nothing; on hard tasks the average dilutes a fragile specialist below even the best single parent and routing wins by a wide margin (Fig. 6A–B).

The generative payoff is the *Fisher–Muller effect* (33, 34): recombination assembles, in one offspring, complementary variants that arose in different lineages, producing a genotype fitter than any parent. In the multi-locus model, sexual merging of decorrelated specialists climbs to the global optimum, a genotype no parent held, while the best single parent and the blended average both plateau below (Fig. 2). In real language models the signature replicates under seed replication: merges of three LoRA (35) specialists beat every parent overall (decisively at 7B: 0.87 vs 0.77), and on the sharper worst-family metric the merged models are the only ones competent everywhere, in every seed (Fig. 6A).

Sex has risks and, for AI, an unfair advantage, both quantified on rugged (epistatic) NK landscapes (36) (Fig. 3). When skills are entangled, blind recombination produces offspring *below* their parents (outbreeding depression), worsening with ruggedness, and the optimal recombination rate shrinks as entanglement grows. But an engineered population can do what biology cannot: recombine unbounded parents, choose complementary mates, and *screen many candidate offspring against a verifier before keeping one*. This directed sex converts the outbreeding catastrophe into a reliable gain in the model (tracking or exceeding the best parent at every ruggedness) and replicates as a sign in language models: bred-and-screened merges beat the a-priori blend in every seed on headroom tasks, including one seed where the blend failed catastrophically and selection was immune (Fig. 6A). Finally, population *structure* is itself a knob: sweeping the mate-pool breadth from monogamous (local) to promiscuous (pan-mictic) against ruggedness, wide mixing maximises the population mean while monotonically destroying diversity, and the best *champion* shifts from wide breadth on smooth landscapes to intermediate breadth on rugged ones (Fig. 3C), the mating-system phenomenon known to structured-population search, mapped onto merging populations.

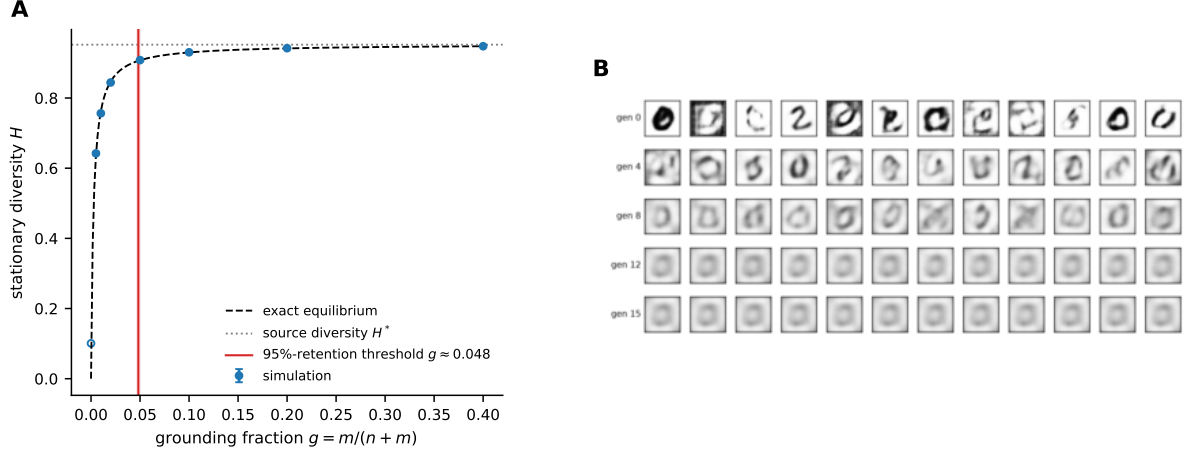


Figure 1: Grounding is immigration. (A) Stationary diversity against the grounding fraction in the minimal inheritance model: simulation (points, 95% CI) matches the exact immigration–drift equilibrium (dashed). The equilibrium is smooth in  $g$ ;  $g \approx 0.05$  marks the operational threshold retaining 95% of source diversity in this setting (red line, bootstrap CI shaded); the hollow point at  $g = 0$  is a finite-time value (the true equilibrium is zero). (B) The same signs on real images: samples from a convolutional VAE retrained each generation on its own output (rows: generations 0–15 of an ungrounded lineage) collapse toward a single blurred mode; 10% grounding holds all thirty modes (quantified in SI).

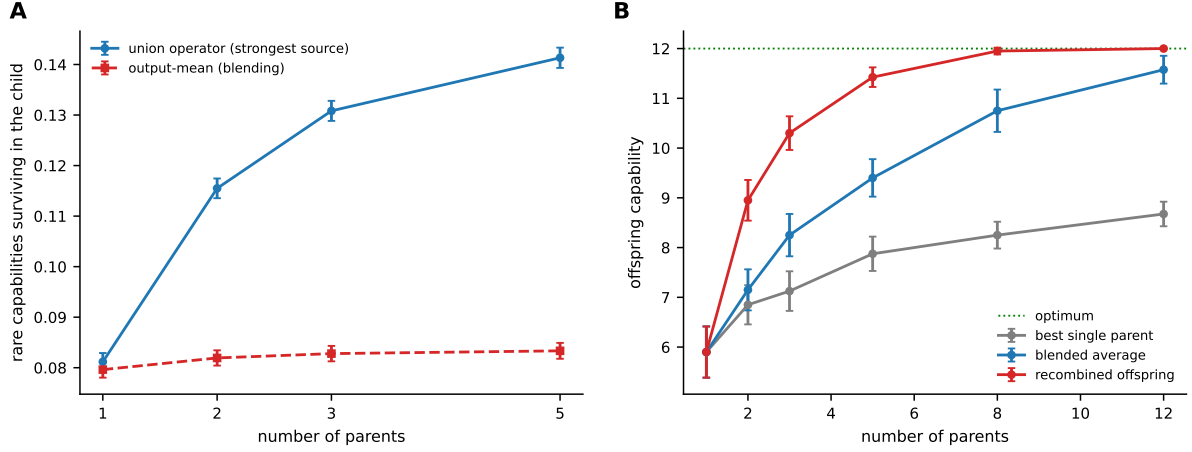


Figure 2: Recombination in the minimal model: blending inheritance and the Fisher–Muller effect. (A) Expected rare-capability survival in a child refit from  $K$  uncorrelated parents: the output-mean (blending) stays at the single-parent level — the first-order cancellation — while the union operator (strongest source per item, renormalised, oracle-identified) rises with parent count. (B) Multi-locus recombination of decorrelated specialists produces offspring fitter than any parent, approaching the optimum as parents are added; the best single parent and the blended average plateau below (mean  $\pm$  95% CI).

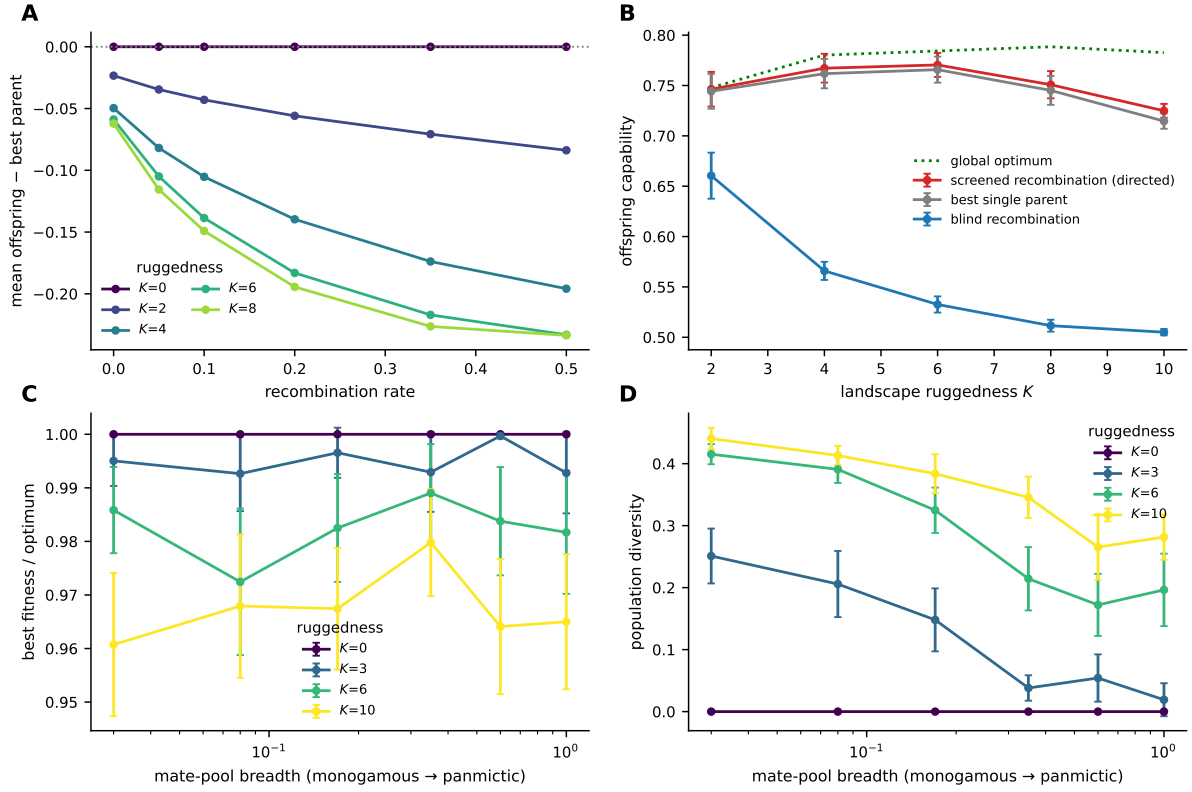


Figure 3: Rugged (epistatic) landscapes: risk, remedy, and population structure. (A) Outbreeding depression: the mean offspring of blindly recombined specialist parents falls below the best parent, more steeply the more rugged the landscape (NK ruggedness  $K$ ) and the higher the recombination rate. (B) Screening candidate offspring against a verifier (directed recombination) restores the gain at every ruggedness where blind recombination fails. (C) Mating structure: the best champion arises at wide mate-pool breadth on smooth landscapes and at intermediate breadth on rugged ones. (D) Wide breadth monotonically erodes population diversity at every ruggedness (mean  $\pm$  95% CI, 20 replicates).

## The society: grounding, recombination, and diversity make complementary contributions

Composing the operators (Fig. 4) requires one definitional distinction first. In the inheritance model, grounding is *grounded inheritance*: external samples added to the reproduction process (the data channel). In the society model, grounding is *grounded evaluation*: selection weights true fitness against conformity to the population’s own consensus,  $g \cdot \text{true-fitness} + (1-g) \cdot \text{conformity}$ , the analogue of scoring models by the crowd’s approval (the fitness channel). These are related design ideas, since both couple the lineage to a non-drifting external signal, but they are different operators, and I name them separately. In the tested society (a finite agent population on a rugged NK landscape), a four-arm ablation separates the failure modes: the full system (grounded evaluation + directed recombination + diversity-preserving selection (37)) climbs to near the global optimum while keeping its specialists; removing grounded evaluation converges the population confidently on an unfit consensus (self-consumption); removing recombination strands it on local optima; removing diversity converges it prematurely to a worse answer. Each removal fails differently; the three implementations make complementary contributions *under the tested conditions*; general joint necessity is not established (alternative mutation, restart, archive, or selection schemes could alter the picture). At language-model scale this composed loop remains unbuilt; it is the paper’s largest stated gap.

## The limit of sex: model speciation

Recombination presupposes compatible parents. In biology, lineages pushed far enough apart become separate species (*reproductive isolation*) through Bateson–Dobzhansky–Muller incompatibilities (38, 39): changes harmless on their own background but deleterious in combination. A merged model is exactly the exposed hybrid. I built the analytic model (Fig. 5A): hybrid fitness tracks the parents while compatible, then peels off and crashes below the ancestor; the isolation cliff arrives earlier the denser the incompatibilities; and the incompatibility *count* snowballs quadratically with divergence (39). Note that a super-linear count does not by itself entail a sharp performance cliff without the count-to-effect-size link, which the analytic model supplies under its assumptions and any neural test must establish separately.

In trained networks, the claim must survive a known alternative: merge barriers between independently trained networks are famously *coordinate artefacts*, removable by re-aligning hidden units (40); richer symmetry groups remove more (41), with known failures beyond the shared-data regime (42). I therefore aligned under the composition of permutation matching and exact per-unit rescaling (the unit symmetry group of plain ReLU MLPs, as the search space) and decomposed the barrier (Fig. 5 C and D): two networks trained from different initialisations on the *same* task have a barrier that this alignment removes essentially entirely (residual  $\approx 0.001$ , the aligned merge performing at parent level): coordinate, not functional; two networks trained on *conflicting* label maps have a barrier the same alignment leaves largely unchanged ( $0.502 \rightarrow 0.497$ ), with the merged model functionally dead. The tested alignment removes the same-task barrier but leaves the conflict-associated barrier intact, supporting a functional-conflict interpretation without proving optimal alignment: exact recovery of a permuted-and-rescaled copy validates a special case, so the removable share is a lower bound and the residual an upper bound. Sweeping conflict traces the cliff as hybrid fitness,  $0.97 \rightarrow 0.03$ . The conflict floor itself is information-theoretic (no single model can satisfy contradictory conventions; SI Appendix, Proposition S2), with the framework’s role being the *structure around it*: which divergences generate conflict, and what moves the cliff.

The pre-registered *emergent test* constrains the claim most: true BDM incompatibilities are emergent (each lineage’s changes harmless alone), so I let children diverge with *no conflicting signal anywhere*, using complementary class specialists and divergent input conventions, to

6.4 $\times$  the base training. No isolation emerged (residual 0.000 throughout); instead the merge *rescued* the two catastrophically-forgetting specialists (parents  $\approx 0.50$ , merge  $\approx 0.955$ , a sustained Fisher–Muller rescue). The same double result appears at the language-model tier (Fig. 5 E and F): conflicting conventions produce *function-specific* hybrid breakdown (the merge scores below both parents on the conflicted function, while a budget-controlled design shows the disjoint skills merge unharmed), and over-training disjoint specialists 1 $\rightarrow$ 12 epochs (cf. the merging literature’s expert-duration effect; 43) produces no isolation at all — the merge improves. Across every tier tested, isolation had to be provoked by functional conflict; specialisation alone did not speciate — a bound on the analogy that sharpens the design rule: what breaks merging is conflicting conventions on shared circuitry, not divergence per se.

### A controlled predictive test: functional conflict, measured pre-merge, predicts merge damage

The framework’s prediction-level claim was put to a designed test (Fig. 6C). Thirty-nine parent pairs (13 conditions  $\times$  3 seeds; rows are not independent — parents share task-data seeds across conditions, so inference is condition-clustered, and because shared seeds also couple rows *across* conditions I report per-seed and leave-one-seed-out sensitivity alongside) span three axes decorrelated by construction: *conflict* (contradictory conventions on shared prompts, private budgets fixed), *compatible overlap* (the same shared prompts under the same convention — overlap and volume without conflict), and *duration* (weight divergence with zero conflict). Before merging, six predictors are computed: *confidence-weighted functional conflict* (bilateral confident disagreement on probes drawn blind to where conflict lives — a proposed proxy for merge-relevant interactions, motivated by the observation that raw disagreement counts harmless complementation, one parent merely ignorant, as conflict), raw disagreement, gradient alignment at the shared base (44), LoRA-delta cosine and distance, and a cross-task performance baseline. The pre-registered outcome is the merge penalty against oracle parent potential (the hybrid-load analogue), also reported against best- and mean-parent references because the predictor ordering is sensitive to that choice.

Across this controlled grid, pre-merge functional disagreement predicted merge penalties (clustered bootstrap CIs excluding zero; held-out leave-one-condition-out  $\rho \approx 0.35$ – $0.40$ ), whereas LoRA-delta cosine and L2 showed no statistically detectable association; gradient alignment carried intermediate signal. Head-to-head predictor differences are not individually significant at this sample size; only these baselines were tested; and with three seeds, uncertainty about seed generalisation remains substantial — though the seed sensitivity favours the functional measures (per-seed  $\rho$  stable at  $+0.37$  to  $+0.53$  in each seed alone, geometry  $\approx 0$  in every seed, gradient alignment seed-unstable at  $-0.11$  to  $-0.55$ ). Two further results bound the claim: the initial two-axis grid’s best predictor was delta-cosine ( $\rho = +0.60$ ) — an overlap artefact that the compatible-overlap control was added to expose, and did (collapse to  $+0.03$ ); and the pre-registered internal prediction that confidence weighting would beat raw disagreement *failed* (they are statistically indistinguishable as rank predictors), so the present evidence favours functional disagreement generally, not the DMI-specific refinement. The framework motivated the measurement and the controls; their success does not validate the specifically population-genetic mechanism. Whether the prediction improves a budget-matched operator choice, and whether it generalises to unfamiliar conflict structures and real task pairs, are the experiment’s open front.

**Table 2.** Headline quantitative results with sample sizes, uncertainty, and outcome definitions (full per-experiment tables and falsifier status in SI Appendix and per-experiment documentation).

| Result                    | Setting / n                                                | Outcome definition                                                                  | Headline                                                                                                       |
|---------------------------|------------------------------------------------------------|-------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| Closed-form validation    | Analytic tier; standing tests                              | Simulated vs closed-form H-decay, immigration equilibrium, multi-teacher union      | Agreement < 0.5%                                                                                               |
| Grounding retention       | Minimal model; 18+ replicates per point                    | Fraction of equilibrium diversity retained at grounding g (operational threshold)   | $g \approx 0.05$ retained $\geq 95\%$ (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), fixed tests; single 7B run                 | Merged vs best-specialist accuracy (overall; worst family)                          | Ties $0.647 \pm 0.027$ vs $0.592 \pm 0.009$ ; 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→12 epochs | Residual barrier; merged vs parent accuracy                                         | 0.000 everywhere; merge rescues parents ( $\approx 0.955$ vs $\approx 0.50$ )                                  |
| Predictive test           | 13 conditions $\times$ 3 seeds (0.5B)                      | Merge penalty vs oracle parent potential (pre-registered; $\pm$ : clustered 95% CI) | Functional $\rho$ $+0.45/+0.46$ , CI excl. 0; LOCO $\rho \approx 0.4$ ; geometry n.s.; paired differences n.s. |

## Discussion

**Design rules.** As engineering guidance, the results reduce to rules that an operator of a model population can apply, answering the four decisions posed in the Introduction. *Ground every generation* in verified reality — a few percent retained most diversity in the tested settings — but price the rarest capabilities individually (observation probability  $1 - e^{-m \cdot p}$  per batch under unstratified sampling), consider targeted sampling for the deep tail, and use recombination to recover rare capabilities still retained across complementary parents. *Merge, don’t blend, when there is headroom:* keep specialists intact and route, or breed-and-screen candidate merges, whenever the naive average is far from ceiling; plain averaging is adequate only where a strong base has already composed the skills. *Match the operator to entanglement:* merge freely when skills are additive; sparingly, with offspring selection, when they entangle; and expect the champion-optimal mating breadth to narrow as landscapes roughen. *Preserve diversity as a first-class objective*, because selection can only preserve variety that exists, and in the tested society its removal produced a distinct failure mode. *Before merging, measure functional conflict* — cheap, pre-merge, and in the controlled setting predictive where the tested weight-distance baselines were not; and *do not treat divergence or specialisation alone as evidence of incompatibility* — in every regime tested here, what broke merging was conflicting conventions on shared circuitry, which is the thing to detect.

**Continual learning at the population scale.** Within a single network, the discipline’s remedies for forgetting are this framework’s operators writ small. Rehearsal and replay of stored data (26, 27) is grounded inheritance within one lineage, and the replay fractions the field settled on empirically, on the order of 1% for instruction tuning (45) and 5% to 25% by distribution-shift strength in continual pretraining (46), sit where the minimal model’s operational threshold

lies. *Pseudo-rehearsal*, the replay of a network’s own generated samples, proposed as a cure in 1995 (47) and revived as generative replay (48), is precisely the ungrounded null studied here: immigration from a drifting source, benign for one hop, compounding over generations, with verifier-filtering (29, 49) converting it back into grounding. Parameter isolation (50), including frozen-base adapters, which forget far less (51), is engineered decorrelation; complementary-learning-systems consolidation (52–54) is the periodic adapter-into-base merge; the recent turn to merging as a continual-learning mechanism (55–58) applies recombination within one lineage over time, where this paper applies it across lineages; and the observation that rare examples and long-tail knowledge are forgotten first (59–61) is tail extinction seen one model at a time. The mechanisms differ (forgetting is largely deterministic interference, collapse is sampling drift) but the victims and the remedies coincide, and to my knowledge no prior work carries population-genetic formalism into continual learning. Read into that field, the results offer: (i) an equilibrium theory for the replay ratio, with the sharper prediction that the required fraction is set by the rarest capability one refuses to lose (the  $1 - e^{-m \cdot p}$  law) rather than by average loss, testable against published replay sweeps; (ii) a *failure theory for generative replay*: self-generated rehearsal is safe for short horizons and compounds into collapse across generations unless verifier-filtered back into grounding (29, 47–49); (iii) *pre-merge interference prediction with a mechanism*: where the current state of the art fits regressions over candidate metrics (44), the functional-conflict measure arrives at a convergent signal from principle and comes with an operator prescription — when conflict is high, do not average; route or breed-and-screen; (iv) a candidate *decision rule for the consolidate-versus-stay-modular question* that currently splits the field’s practice (keep adapters separate vs merge them; 54–58): union-preserving operators where headroom exists, fusion where the base composes, consolidation as the slow-store step; and (v) *tail monitoring as the leading indicator*: continual-learning evaluation that averages over capabilities hides exactly the losses that drift theory says come first and, past a threshold, become irreversible. On that last point I note the standing objection that apparent forgetting can be skewed task-inference over latent capability rather than erasure (62); the irreversibility results here concern oracle-measured behavioural distributions, and distinguishing latent from extinct capability at language-model scale is an open experiment whose outcome would be decisive for both readings.

**What is borrowed and what is new.** The collapse-as-drift diagnosis is established prior work (21–25); so are the empirical facts that merges can beat parents, that decorrelated parents merge better, and that naive averaging loses to interference-aware or routed merges (4, 63, 64), that model populations can climb (5, 8–10), and that merge success admits ML-native predictors (44, 65), correlational where this framework supplies mechanism; the reading of sex as an algorithm for mixability in the theory of computation (66) anticipated the transfer before model merging existed. New here is the framework-level synthesis — inheritance, diversity, and compatibility as managed quantities — together with: the conservation law for blending inheritance and its operator boundaries; the per-item grounding floor; the society ablation with its complementary failure modes; model speciation as a named, tested question, with the coordinate-versus-functional decomposition under permutation-and-rescaling alignment and the emergent null that bounds it; and the controlled predictive test with its controls. I claim the framework generated these measurements and experiments; I do not claim that their outcomes validate a uniquely population-genetic mechanism, and one refinement it proposed was not supported.

**Limits and open problems.** The demonstrations are deliberately small: exact where small is a virtue, sign-level and seed-replicated at the language-model tier, on constructed task families with a trivially separable router and one model lineage (Qwen, 0.5B–7B). The composed society has not been built at language-model scale. The predictive test’s next bars, in order of value: generalisation to *unfamiliar* conflict structures and real task pairs; a demonstrably better *budget-matched* merging decision; then scale replication. Beyond engineering, the framework’s

hardest open problem is the fitness function itself: selection optimises what is measured, and for knowledge systems the persuasive and the true compete — grounding against a reality that can refuse is the only anchor I trust, and institutionalising that anchor (verification, replication, and challenge among models) is the society-level problem this paper poses but does not solve. What biology receives in return is a new model system: populations of learners where every genotype, environment, and mating decision is observable and manipulable — where the evolution of sex can be studied with interventions (unbounded parents, offspring preview, directed mating) that no living system permits.

**Creative diversity.** Collapse is not confined to facts and skills. Homogenisation of *style* is already measurable: models trained on model output lose lexical and syntactic diversity across generations (67), writing produced with model assistance is individually better but collectively less diverse than writing produced without it (68, 69), and the house styles of the large assistants are recognisable enough that their tics serve as signatures. In this framework these are the same phenomenon at a different locus. A voice is a distribution over rare stylistic variants, exactly the tail that drift erases first and that blending inheritance averages into a common register. The remedies transfer unchanged, though they are untested here: grounding on stylistically diverse human sources, decorrelated lineages maintained as distinct voices rather than merged into one, union-preserving recombination over blending, and selection that rewards being different as well as being good. Whether these preserve measured stylistic diversity at scale is an open experiment that the framework specifies.

**Outlook: the evolution of language models.** The Introduction’s premise, that the model ecosystem is an evolving population, is also a forecast about where these results matter next. Language-model development is consolidating around exactly the operators studied here: synthetic-data flywheels (inheritance), merging and routing of specialist fine-tunes (recombination and population structure), verifier-gated data pipelines (grounded selection), and periodic consolidation of adapters into new bases. If coming model generations remain what the tested regimes found, freely recombinable in the absence of conflicting conventions, then the ecosystem evolves as one interbreeding population, and the levers that matter are grounding budgets priced per rare capability and diversity preserved deliberately. If instead long-horizon specialisation at scale begins to produce emergent incompatibility, as the expert-training-duration observations hint (43) and the small-scale null here does not rule out, then lineages will begin to speciate, and the ecosystem’s future is a set of diverging species connected by routing rather than by merging. Which of the two it will be is measurable now, with the pre-merge conflict instruments this paper tested.

## Materials and Methods

**Analytic tier.** Pure NumPy/SciPy Wright–Fisher simulator over  $K$ -item distributions (knowledge as  $\mathbf{p}_t$ ; Zipf-tailed truth  $\mathbf{p}^*$ ; drift–grounding–refit generations), extended with a learning kernel (smoothing/sharpening refit), multi-locus genotypes on additive and Kauffman NK landscapes,  $n$ -parent crossover, and finite-population society loops. All parameters live in per-experiment YAML configs; every run derives all randomness from one master seed (`SeedSequence.spawn`) and is bitwise reproducible; scientific-validation tests assert the closed forms (heterozygosity decay, immigration equilibrium, closed-form union) to  $<0.5\%$  and run in CI with 151 further correctness tests.

**Neural tier.** Trained-network experiments realise the same abstractions with an exact oracle: histogram/RNN/MLP/VAE generators on a synthetic mode universe (the histogram model reduces the harness exactly to the analytic tier — the bridge gate), and a convolutional VAE on MNIST with a frozen CNN oracle (98.5% mode accuracy; confusion matrix recorded as the measurement floor). Speciation experiments fork no-BatchNorm MLPs (784–512–512–10)

from a shared base, weight-average, and measure linear-mode-connectivity error barriers before and after alignment; alignment composes deterministic Git Re-Basin permutation matching with exact per-unit scale canonicalisation (the unit symmetry group of this class, as the alignment search space; control recovery does not establish global optimality), gated by exact recovery of a permuted-and-rescaled copy.

**Language-model tier.** LoRA specialists (rank 16) on procedurally generated task families with an exact-match verifier, on frozen Qwen2.5-Instruct bases (0.5B on one 16 GB GPU; 7B on one L40S). Operators: weight-space merges (soup/TIES via adapter arithmetic), per-input routing, and Dirichlet-sampled offspring populations screened on held-out validation splits. Multi-seed protocols fix the test sets and vary the training seed. The predictive test computes all predictors pre-merge (generation confidence from token log-probabilities; base-model gradient cosines; exact r-space LoRA-delta geometry) and evaluates merges on held-out tests; robust statistics (condition-clustered bootstrap, paired predictor contrasts, leave-one-condition-out prediction, multi-reference outcomes) are produced by a committed script. Statistical, per-seed reproducibility is documented for GPU tiers.

**Data and code availability.** All code, configs, seeds, results artifacts (with content hashes), figures, and a one-command reproduction script will be deposited openly (repository + archived DOI) on publication; every figure in this paper regenerates from committed artifacts without re-simulation.

## References

1. B. Laufer, H. Oderinwale, J. Kleinberg, Anatomy of a machine learning ecosystem: 2 million models on Hugging Face. *arXiv [Preprint]* (2025). <https://doi.org/10.48550/arXiv.2508.06811>.
2. E. Horwitz, A. Shul, Y. Hoshen, Unsupervised model tree heritage recovery. *Int. Conf. Learn. Represent.* (2025). <https://doi.org/10.48550/arXiv.2405.18432>.
3. W. Jiang, et al., PeaTMOSS: A dataset and initial analysis of pre-trained models in open-source software. *Proc. Int. Conf. Min. Softw. Repos.* (2024). <https://doi.org/10.48550/arXiv.2402.0069>.
4. P. Yadav, D. Tam, L. Choshen, C. Raffel, M. Bansal, TIES-Merging: Resolving interference when merging models. *Adv. Neural Inf. Process. Syst.* **36** (2023). <https://doi.org/10.48550/arXiv.2405.18432>.
5. T. Akiba, M. Shing, Y. Tang, Q. Sun, D. Ha, Evolutionary optimization of model merging recipes. *Nat. Mach. Intell.* **7**, 195–204 (2025).
6. C. Goddard, et al., Arcee’s MergeKit: A toolkit for merging large language models. *Proc. Conf. Empir. Methods Nat. Lang. Process. (Industry Track)*, 477–485 (2024). <https://doi.org/10.48550/arXiv.2403.13257>.
7. E. Yang, et al., Model merging in LLMs, MLLMs, and beyond: Methods, theories, applications and opportunities. *arXiv [Preprint]* (2024). <https://doi.org/10.48550/arXiv.2408.07666>.
8. Y. Zhang, et al., Nature-inspired population-based evolution of large language models (GENOME). *arXiv [Preprint]* (2025). <https://doi.org/10.48550/arXiv.2503.01155>.
9. J. Abrantes, et al., Competition and attraction improve model fusion (M2N2). *Proc. Genet. Evol. Comput. Conf.* (2025). <https://doi.org/10.48550/arXiv.2508.16204>.
10. V. Subramaniam, et al., Multiagent finetuning: Self-improvement with diverse reasoning chains. *arXiv [Preprint]* (2025). <https://doi.org/10.48550/arXiv.2501.05707>.

11. NVIDIA (B. Adler, et al.), Nemotron-4 340B technical report. arXiv [Preprint] (2024). <https://doi.org/10.48550/arXiv.2406.11704>.
12. M. Abdin, et al., Phi-4 technical report. arXiv [Preprint] (2024). <https://doi.org/10.48550/arXiv.2412.089>.
13. Y. Wang, et al., Self-Instruct: Aligning language models with self-generated instructions. *Proc. Annu. Meet. Assoc. Comput. Linguist.* (2023). <https://doi.org/10.48550/arXiv.2212.10560>.
14. B. Thompson, et al., A shocking amount of the web is machine translated: Insights from multi-way parallelism. *Findings Assoc. Comput. Linguist.: ACL* (2024). <https://doi.org/10.48550/arXiv>.
15. W. Liang, et al., Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. *Proc. Int. Conf. Mach. Learn.* (2024). <https://doi.org/10.48550/arXiv.2403.07183>.
16. P. Villalobos, et al., Position: Will we run out of data? Limits of LLM scaling based on human-generated data. *Proc. Int. Conf. Mach. Learn.* (2024). <https://doi.org/10.48550/arXiv.2211.043>.
17. L. Brinkmann, et al., Machine culture. *Nat. Hum. Behav.* **7**, 1855–1868 (2023).
18. J. S. Park, et al., Generative agents: Interactive simulacra of human behavior. *Proc. ACM Symp. User Interface Softw. Technol.* (2023). <https://doi.org/10.48550/arXiv.2304.03442>.
19. T. Guo, et al., Large language model based multi-agents: A survey of progress and challenges. *Proc. Int. Joint Conf. Artif. Intell.* (2024). <https://doi.org/10.48550/arXiv.2402.01680>.
20. N. Tomasev, et al., Virtual agent economies. arXiv [Preprint] (2025). <https://doi.org/10.48550/arXiv.250>.
21. I. Shumailov, et al., AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024).
22. J. P. Crutchfield, S. Whalen, Structural drift: The population dynamics of sequential learning. *PLOS Comput. Biol.* **8**, e1002510 (2012).
23. S. Riis, Drift and selection in LLM text ecosystems. arXiv [Preprint] (2026). <https://doi.org/10.48550/arXiv>.
24. M. Benati, A. Londei, D. Lanzieri, V. Loreto, First-extinction law for resampling processes. arXiv [Preprint] (2025). <https://doi.org/10.48550/arXiv.2509.20101>.
25. Y. Yoon, D. Hu, I. Weissburg, Y. Qin, H. Jeong, Model collapse in the self-consuming chain of diffusion finetuning: A novel perspective from quantitative trait modeling. *Int. Conf. Learn. Represent.* (2025). <https://doi.org/10.48550/arXiv.2407.17493>.
26. M. McCloskey, N. J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem. *Psychol. Learn. Motiv.* **24**, 109–165 (1989).
27. R. M. French, Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* **3**, 128–135 (1999).
28. H. J. Muller, The relation of recombination to mutational advance. *Mutat. Res.* **1**, 2–9 (1964).
29. B. Yi, Q. Liu, Y. Cheng, H. Xu, Escaping model collapse via synthetic data verification. arXiv [Preprint] (2025). <https://doi.org/10.48550/arXiv.2510.16657>.
30. M. Gerstgrasser, et al., Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. *Conf. Lang. Model.* (2024). <https://doi.org/10.48550/arXiv.2404.014>.
31. S. Wright, Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931).

32. J. Pari, S. Jelassi, P. Agrawal, Collective model intelligence requires compatible specialization. *arXiv [Preprint]* (2024). <https://doi.org/10.48550/arXiv.2411.02207>.
33. R. A. Fisher, *The Genetical Theory of Natural Selection* (Clarendon Press, 1930).
34. H. J. Muller, Some genetic aspects of sex. *Am. Nat.* **66**, 118–138 (1932).
35. E. J. Hu, et al., LoRA: Low-rank adaptation of large language models. *Int. Conf. Learn. Represent.* (2022). <https://doi.org/10.48550/arXiv.2106.09685>.
36. S. A. Kauffman, S. Levin, Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.* **128**, 11–45 (1987).
37. J. Lehman, K. O. Stanley, Abandoning objectives: Evolution through the search for novelty alone. *Evol. Comput.* **19**, 189–223 (2011).
38. H. A. Orr, The population genetics of speciation: The evolution of hybrid incompatibilities. *Genetics* **139**, 1805–1813 (1995).
39. H. A. Orr, M. Turelli, The evolution of postzygotic isolation: Accumulating Dobzhansky–Muller incompatibilities. *Evolution* **55**, 1085–1094 (2001).
40. S. K. Ainsworth, J. Hayase, S. Srinivasa, Git Re-Basin: Merging models modulo permutation symmetries. *Int. Conf. Learn. Represent.* (2023). <https://doi.org/10.48550/arXiv.2209.04836>.
41. T. Li, Z. Shen, Scaling linear mode connectivity and merging to billion-parameter pre-trained transformers. *arXiv [Preprint]* (2026). <https://doi.org/10.48550/arXiv.2606.23607>.
42. E. Sharma, D. M. Roy, G. K. Dziugaite, The non-local model merging problem: Permutation symmetries and variance collapse. *arXiv [Preprint]* (2024). <https://doi.org/10.48550/arXiv.2410.12706>.
43. N. Kozodoi, Z. Afolabi, J. Butler, Are we merging the right models? Impact of expert training duration on model merging for LLMs. *arXiv [Preprint]* (2026). <https://doi.org/10.48550/arXiv.2607.12706>.
44. L. Zhou, B. Zhao, R. Yu, E. Rodolà, Demystifying mergeability: Interpretable properties to predict model merging success. *arXiv [Preprint]* (2026). <https://doi.org/10.48550/arXiv.2601.22285>.
45. T. Scialom, T. Chakrabarty, S. Muresan, Fine-tuned language models are continual learners. *Proc. Conf. Empir. Methods Nat. Lang. Process.* (2022). <https://doi.org/10.48550/arXiv.2205.12390>.
46. A. Ibrahim, et al., Simple and scalable strategies to continually pre-train large language models. *Trans. Mach. Learn. Res.* (2024). <https://doi.org/10.48550/arXiv.2403.08763>.
47. A. Robins, Catastrophic forgetting, rehearsal and pseudorehearsal. *Connect. Sci.* **7**, 123–146 (1995).
48. H. Shin, J. K. Lee, J. Kim, J. Kim, Continual learning with deep generative replay. *Adv. Neural Inf. Process. Syst.* **30** (2017). <https://doi.org/10.48550/arXiv.1705.08690>.
49. Y. Feng, et al., Beyond model collapse: Scaling up with synthesized data requires verification. *arXiv [Preprint]* (2024). <https://doi.org/10.48550/arXiv.2406.07515>.
50. A. A. Rusu, et al., Progressive neural networks. *arXiv [Preprint]* (2016). <https://doi.org/10.48550/arXiv.1606.04671>.
51. D. Biderman, et al., LoRA learns less and forgets less. *Trans. Mach. Learn. Res.* (2024). <https://doi.org/10.48550/arXiv.2405.09673>.

52. J. L. McClelland, B. L. McNaughton, R. C. O'Reilly, Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
53. D. Kumaran, D. Hassabis, J. L. McClelland, What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn. Sci.* **20**, 512–534 (2016).
54. J. Schwarz, et al., Progress & Compress: A scalable framework for continual learning. *Proc. Int. Conf. Mach. Learn.* (2018).
55. G. Ilharco, et al., Editing models with task arithmetic. *Int. Conf. Learn. Represent.* (2023). <https://doi.org/10.48550/arXiv.2212.04089>.
56. D. Marczak, B. Twardowski, T. Trzciński, S. Cygert, MagMax: Leveraging model merging for seamless continual learning. *Proc. Eur. Conf. Comput. Vis.* (2024). <https://doi.org/10.48550/arXiv.2407.08612>.
57. A. Alexandrov, et al., Mitigating catastrophic forgetting in language transfer via model merging. *Findings Assoc. Comput. Linguist.: EMNLP* (2024). <https://doi.org/10.48550/arXiv.2407.08612>.
58. S. Dziadzio, et al., How to merge your multimodal models over time? *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (2025). <https://doi.org/10.48550/arXiv.2412.06712>.
59. M. Toneva, et al., An empirical study of example forgetting during deep neural network learning. *Int. Conf. Learn. Represent.* (2019). <https://doi.org/10.48550/arXiv.1812.05159>.
60. N. Kandpal, H. Deng, A. Roberts, E. Wallace, C. Raffel, Large language models struggle to learn long-tail knowledge. *Proc. Int. Conf. Mach. Learn.* (2023). <https://doi.org/10.48550/arXiv.2211.00266>.
61. X. Liu, et al., Long-tailed class incremental learning. *Proc. Eur. Conf. Comput. Vis.* (2022). <https://doi.org/10.48550/arXiv.2210.00266>.
62. S. Kotha, J. M. Springer, A. Raghunathan, Understanding catastrophic forgetting in language models via implicit inference. *Int. Conf. Learn. Represent.* (2024). <https://doi.org/10.48550/arXiv.2407.08612>.
63. L. Yu, B. Yu, H. Yu, F. Huang, Y. Li, Language models are super Mario: Absorbing abilities from homologous models as a free lunch. *Proc. Int. Conf. Mach. Learn.* (2024). <https://doi.org/10.48550/arXiv.2311.03099>.
64. M. Wortsman, et al., Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *Proc. Int. Conf. Mach. Learn.* (2022). <https://doi.org/10.48550/arXiv.2203.05482>.
65. Y. Cao, et al., An empirical study and theoretical explanation on task-level model-merging collapse. *arXiv [Preprint]* (2026). <https://doi.org/10.48550/arXiv.2603.09463>.
66. A. Livnat, C. Papadimitriou, Sex as an algorithm: The theory of evolution under the lens of computation. *Commun. ACM* **59**, 84–93 (2016).
67. Y. Guo, G. Shang, M. Vazirgiannis, C. Clavel, The curious decline of linguistic diversity: Training language models on synthetic text. *Findings Assoc. Comput. Linguist.: NAACL* (2024). <https://doi.org/10.48550/arXiv.2311.09807>.
68. V. Padmakumar, H. He, Does writing with language models reduce content diversity? *Int. Conf. Learn. Represent.* (2024). <https://doi.org/10.48550/arXiv.2309.05196>.
69. A. R. Doshi, O. P. Hauser, Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Sci. Adv.* **10**, eadn5290 (2024).

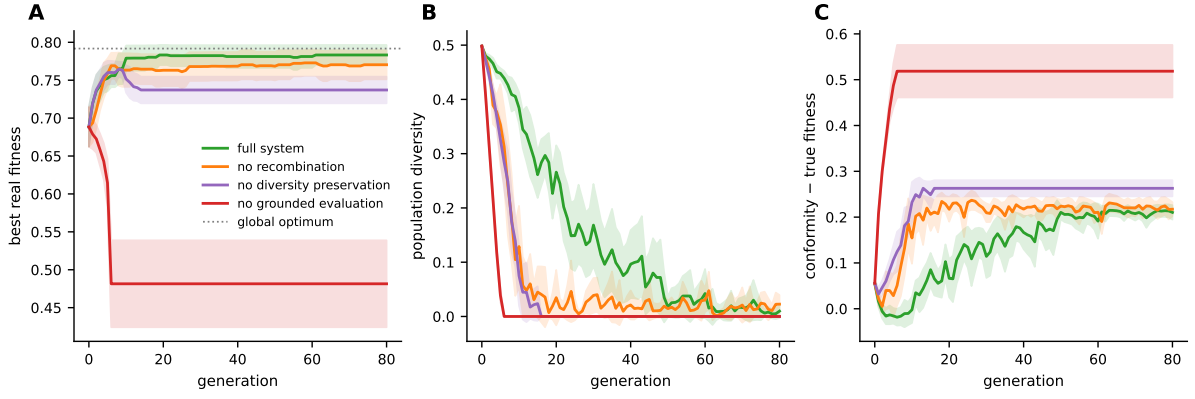


Figure 4: The tested society: grounded evaluation, recombination, and diversity preservation make complementary contributions. A finite agent population on a rugged NK landscape; selection weights true fitness against conformity to the population consensus. (A) Best real fitness: the full system approaches the global optimum; removing grounded evaluation collapses the population onto a confident, unfit consensus; removing recombination or diversity preservation strands it lower. (B) Population diversity. (C) The self-consumption signature: conformity minus true fitness (mean  $\pm$  95% CI, 12 replicates).

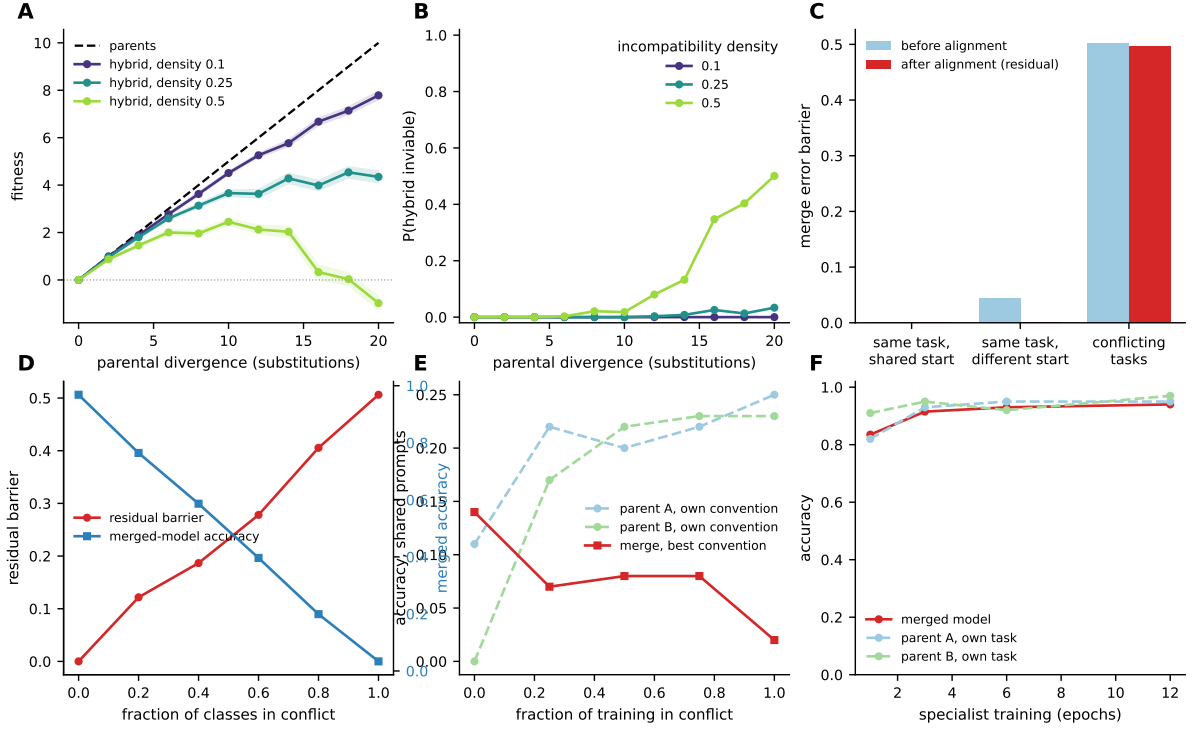


Figure 5: Model speciation at three tiers. (A) Analytic model: hybrid fitness tracks the parents while lineages are compatible, then falls to inviability; the denser the incompatibilities, the earlier the fall. (B) The isolation cliff: probability of hybrid inviability against divergence, by incompatibility density. (C) Trained networks: the merge error barrier between two MLPs before and after permutation-and-rescaling alignment — the same-task/different-start barrier is a coordinate artefact (removed by alignment); the conflicting-task barrier is left essentially unchanged. (D) Sweeping the fraction of conflicting classes: the residual barrier rises while merged-model accuracy falls from 0.97 to 0.03. (E) Language models (0.5B LoRA children of a shared base): on shared ambiguous prompts each parent performs under its own convention while the merged model falls below both — function-specific hybrid breakdown. (F) Divergence without conflict: over-training disjoint specialists from 1 to 12 epochs produces no isolation; the merged model tracks or exceeds the parents throughout.

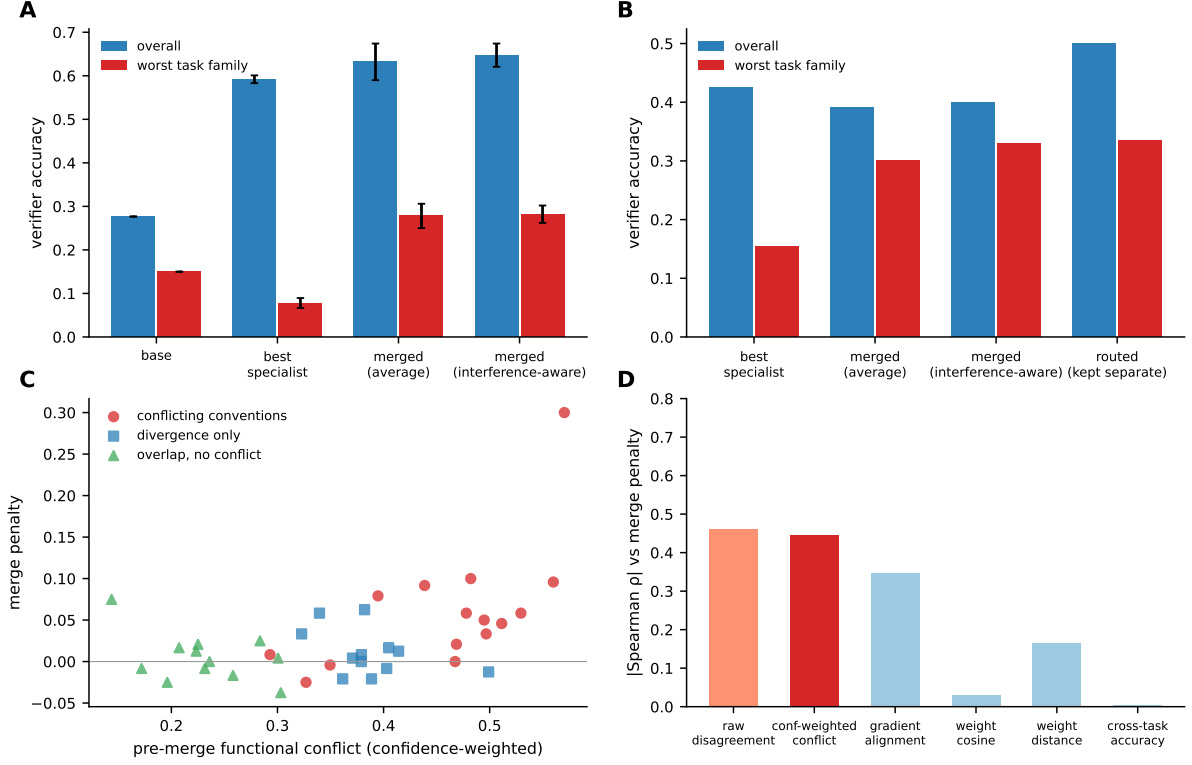


Figure 6: The language-model tier. (A) Seed-replicated merging (0.5B, five seeds, fixed test sets; mean  $\pm$  95% CI): merged specialists exceed the best single specialist overall, and only merged models are competent on every task family. (B) Hard, unsaturated tasks at 7B (single run): the weight-average dilutes a fragile specialist below the best single parent; routing among intact specialists preserves it. (C) The controlled predictive test (13 conditions  $\times$  3 seeds): pre-merge confidence-weighted functional conflict against merge penalty, coloured by grid axis — penalty concentrates on the conflict axis. (D) Predictor comparison, |Spearman  $\rho$ | against merge penalty over the full grid: functional measures carry signal, the tested weight-geometry baselines do not; paired differences between predictors are not individually significant.