

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 is shifting from single, frozen models to populations of models that specialise, are retrained on each other’s output, and are combined (“merged”) into new models. Trained on their own output, model lineages degenerate — a process already recognised as the mathematics of genetic drift. This paper imports the other half of population genetics: the biology of sexual reproduction. It treats model merging as recombination, real data as immigration, and merge failure as reproductive isolation, and tests each correspondence in simulations, small neural networks, and language models. The framework recasts machine learning’s oldest problem — continual learning without forgetting — at the population scale, and yields design rules: when to average models, when to keep them separate, how much real data suffices (a theory for the field’s empirical replay fractions), and a controlled small-model test in which pre-merge functional disagreement predicted merge damage, motivating further comparison with weight-space measures.

Abstract

AI development increasingly resembles a population process: models are specialised, retrained on model output, and recombined by weight merging, with an openly evolutionary vocabulary but little use of evolutionary theory. Here we treat multigenerational model populations as systems whose inheritance, diversity, and compatibility must be managed, and transfer the quantitative apparatus of the evolution of sex. Its entry point — training on model output is genetic drift, and model collapse is its signature — we developed independently, and parallel work has now formalised the same diagnosis from several directions, a convergence we read as evidence for the frame rather than as a shared discovery to be divided. In a minimal inheritance model that is exactly Wright–Fisher — and measurably Wright–Fisher-plus-bias in trained networks — we derive and test the remedies: grounding as immigration, where a real-data fraction far below one retained most equilibrium diversity in the tested settings, with a per-capability observation floor that makes the rarest knowledge expensive under unstratified sampling; recombination, where refitting to the mean of parents’ output distributions cancels the multi-parent gain to first order in the rare-item regime while union-preserving operators realise it; the Fisher–Muller effect, with merged language-model specialists exceeding every parent in replicated experiments; outbreeding depression on rugged task landscapes, converted into reliable gains by directed, offspring-screened recombination; and population structure, where the optimal mating breadth shrinks as skills entangle. Sex has a limit: we introduce model speciation — merge failure as

reproductive isolation — and show in trained networks that a merge barrier remaining after permutation-and-rescaling alignment tracks functional conflict, that isolation did not emerge from compatible specialisation, and, in a controlled predictive test, that pre-merge functional disagreement predicted merge damage while the tested weight-geometry baselines showed no detectable association. We state precisely what is exact, what is measured, and what remains hypothesis.

Introduction

Machine learning has quietly become a population-scale phenomenon. Public repositories now 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 sprawling family trees whose lineage structure, inherited traits, and mutation dynamics are already being mapped with explicitly phylogenetic methods (31–33). Reproduction in this population is no longer metaphorical. **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 (1, 2, 37, 38), and the engineering literature describes it in openly evolutionary vocabulary: “crossover,” “mutation,” “mate choice,” populations of merging models that climb benchmarks (2–5).

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 (43, 44) — self-generated instruction data seeds whole lineages of descendants (5), a large and growing share of the public web is machine-generated or machine-translated text (35, 36), and the stock of human text is projected to be exhausted by frontier training within this decade (34). Meanwhile persistent multi-agent systems and emerging agent economies put many interacting models into sustained contact (39–42). A population whose members inherit from one another, recombine, and retransmit under these conditions is an evolving population in the technical sense, whatever one thinks of the metaphors. The organising claim of this paper is that the vocabulary deserves its mathematics: **multigenerational model populations are systems whose inheritance, diversity, and compatibility must be managed — not merely collections of models to optimise — and the branch of biology that studies exactly this problem, the population genetics of the evolution of sex, transfers as a quantitative framework.**

The frame’s entry point is the diagnosis. 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 (6). That this is the mathematics of **genetic drift** in a finite population is a conclusion we reached independently in building the present framework, and one that has been derived in parallel from several other directions (7–9), including a closed-form first-extinction law placing collapse onset at the Wright–Fisher first-extinction time (8) — and that was anticipated, before deep learning, in an analysis of sequential inference chains as generalised genetic drift (63). We cite these works for priority of publication on the diagnosis and read the convergence — independent arrivals at the same population-genetic account by different routes and in different decades — as corroboration that the frame is the natural one. What none of that parallel work develops, and what this paper is about, is the structure the diagnosis opens: the full arc from drift through its remedies (immigration, recombination, selection, population structure) to its limit (reproductive isolation), carried as one framework from closed forms to trained networks to language models.

Seen from machine learning’s own history, the problem this frame addresses is the field’s

oldest — **continual learning** — reappearing one level up. Within a single network, sequential learning overwrites prior knowledge (catastrophic forgetting; 45, 46), and the discipline’s remedies are, one by one, the population operators of this paper in single-model form: **rehearsal and replay** of past data is grounding’s within-lineage counterpart, and the field’s empirically settled replay fractions — on the order of 1% for instruction tuning (53), 5% for weak and 25% for strong distribution shift in continual pretraining (52) — sit exactly where the minimal model’s operational grounding threshold lies, a correspondence for which the framework supplies the missing theory (equilibrium diversity, and a per-capability survival law). **Pseudo-rehearsal** — replaying the network’s own generated samples, proposed as a cure in 1995 (47) and revived as generative replay (48) — is precisely this paper’s ungrounded null: immigration from a drifting source, benign for one hop and compounding into collapse over generations, with verifier-filtering (12, 62) as what converts it back into grounding. **Parameter isolation** (65, and frozen-base adapters, which forget far less; 54) is the engineered decorrelation our specialists use; **complementary-learning-systems consolidation** (49–51) is our 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 we apply it across lineages; and the observation that **rare examples and long-tail knowledge are forgotten first** (59–61) is tail-allele extinction observed one model at a time. One distinction is kept explicit throughout: catastrophic forgetting is largely deterministic interference from shifted training, whereas collapse is stochastic sampling drift — the phenomena share their victims (the rare) and their remedies, not their mechanism. To our knowledge, no prior work carries population-genetic formalism into continual learning itself; that bridge — replay as immigration with a survival law, merging as recombination with a compatibility criterion, consolidation as the slow store of a two-speed memory — is where this framework may matter most.

We are explicit about what kind of contribution each claim is, distinguishing **interpretation** (an existing result understood in population-genetic terms), **explanation** (the transferred mechanism accounts for observations existing accounts leave open), and **prediction** (the framework forecasts an unmeasured outcome). The paper is strongest on the first; makes concrete progress on the second — separating merge failures that are coordinate artefacts from those that are functional; and reports a first, bounded step on the third — a controlled predictive test in which pre-merge functional-disagreement measures, chosen by the framework, predicted merge damage on a constructed task grid while the tested weight-geometry baselines showed no detectable association.

The correspondences we develop, summarised in Table 1: single-teacher retraining is **asexual reproduction**, and the irreversible arm of its decay shares the defining consequence of **Muller’s ratchet** (10) — once every copy of a rare capability is gone from all parents and sources, no recombination can rebuild it, which is why remedies must act before fixation-by-loss (a consequence- level correspondence: the minimal model lacks the ratchet’s recurrent deleterious-mutation mechanism, so irreversible loss alone does not identify that specific mechanism). Injecting verified real data is **immigration** from a non-drifting source (11–13). Model merging is **recombination**, and its celebrated payoff — a merged model exceeding every parent — is the **Fisher–Muller effect** (14, 15). Merging entangled skills courts **outbreeding depression**; screening many candidate merges is engineered recombination with unusually flexible parent choice and pre-deployment screening (we use the shorthand **directed sex**); restricting who merges with whom is **population structure**. And merging’s hard limit — models too diverged in function to combine — is **reproductive isolation**, for which the Bateson–Dobzhansky–Muller theory of incompatibilities (16, 17) supplies the structure. The nearest precursor to this programme reads sex as an algorithm for mixability in the theory of computation (18), pre-dating model merging; the model-merging literature itself has strong empirical operators (1, 19, 20) and emerging merge-success predictors (21, 22), to which our delta is mechanism: *when and why* failure is coordinate versus functional, and what moves the boundary.

We support the framework at three tiers of evidence, in ascending realism and descending exactness: a **minimal analytic model** validated against closed forms to a fraction of a percent; **small trained networks** (MLPs, recurrent networks, an MNIST image generator) where the operators are measured in real weights; and **language models** (LoRA-specialised Qwen models, 0.5B locally and 7B on a compute cluster) where the claims are tested as signs under seed replication. Negative results are reported with the same prominence as confirmations; they include the failure of an internal pre-registered prediction, a null on emergent speciation that bounds the analogy, and the sensitivity analyses on the predictive test.

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 n samples from the parent’s distribution, optionally mix in m verified real samples (“grounding”, $g = m/(n+m)$), and refit the child*. In this minimal inheritance model the resampling step **is** the Wright–Fisher process — the same equations, which we exploit as an engineering gate: our 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 we 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 we 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).

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.

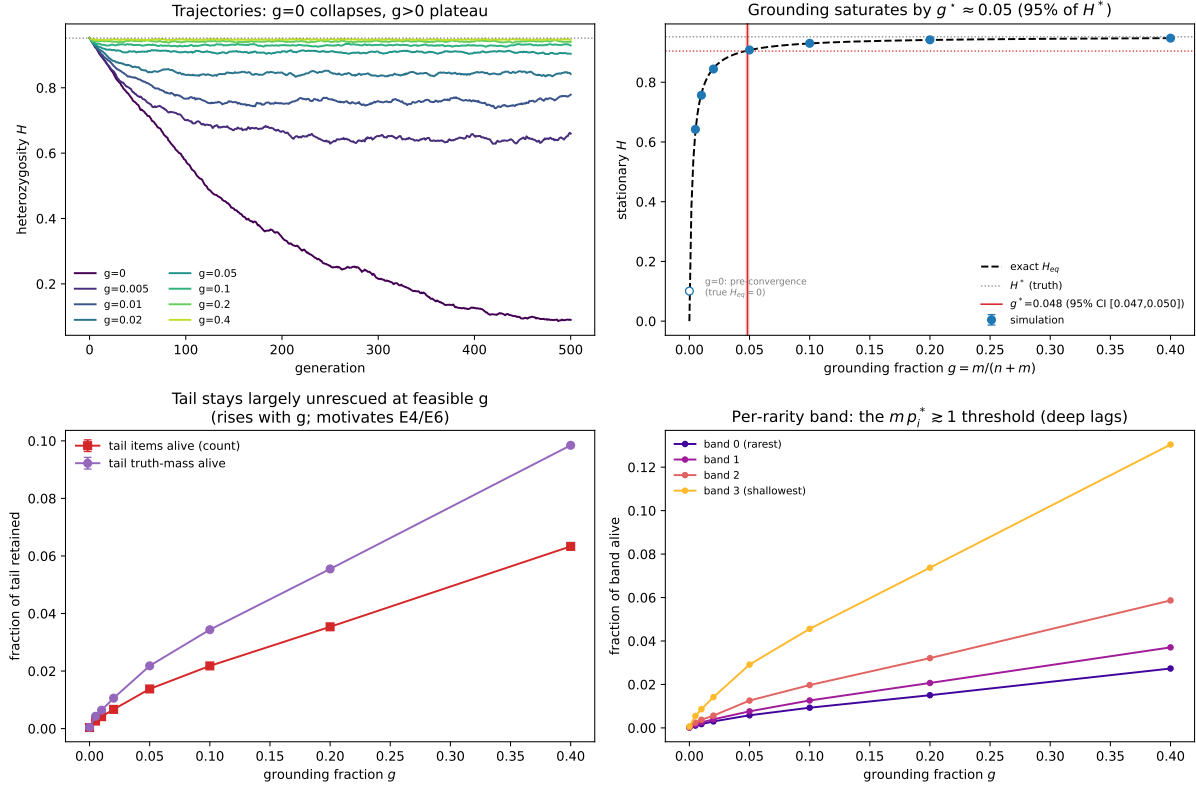
Population genetics	Model populations	Support
Genetic drift in a finite population	Training on finite samples of model output	Exact (minimal model); signs in trained nets; diagnosis conceded to prior work
Immigration from a fixed source	Grounding with verified real data	Exact equilibrium; signs in RNN/MLP/VAE/MNIST
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, and the equilibrium diversity has a closed form our 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 we 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). The engineering point survives the definition: verified real data is 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 we 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).

E2 — a critical grounding ratio $g^* \ll 1$ rescues diversity; the deep tail needs recombination



Dry MNIST lineage collapse: varied digits (top) -> a single mode (bottom)

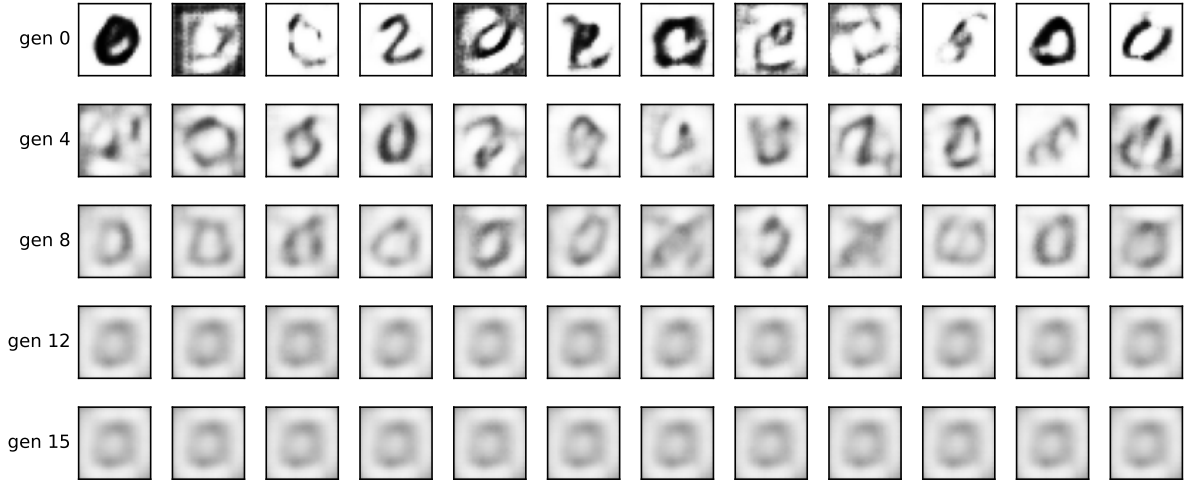


Figure 1: Collapse is drift; grounding is immigration. (A, top) The grounding phase response in the minimal model: a critical real-data fraction $g^* \approx 0.05$ retains most diversity indefinitely, while tail survival obeys the per-item floor $mp \gtrsim 1$. (B, bottom) The same signs on real images: a convolutional VAE retrained each generation on its own output collapses to a single blurred mode (rows: generations), while $\sim 10\%$ grounding holds all thirty class $\times$ style modes.

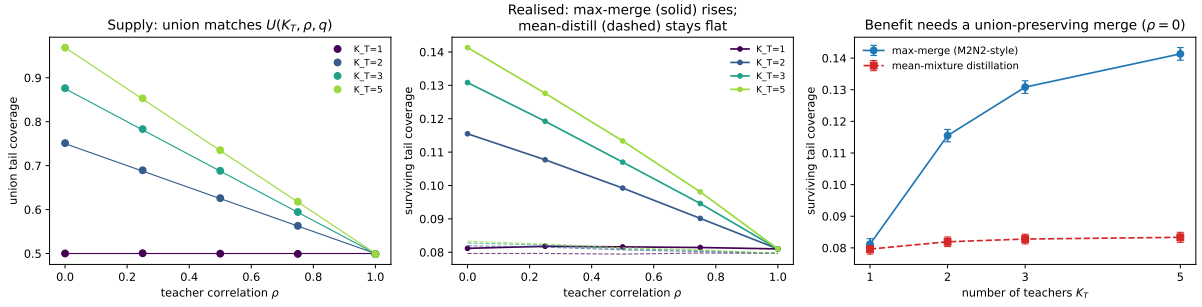
Recombination: a conservation law, its operators, and offspring that exceed every parent

Merging is where the evolution-of-sex apparatus pays for itself, beginning with a result about the obvious operator, stated with its assumptions. **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** (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**: 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 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 (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 (panmictic) 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.

E4 — recombination supplies the tail; only a union-preserving merge realises it in the pupil



E8 — the vertical claim: an offspring recombined from many decorrelated parents is fitter than any parent (Fisher-Muller; no two-parent limit)

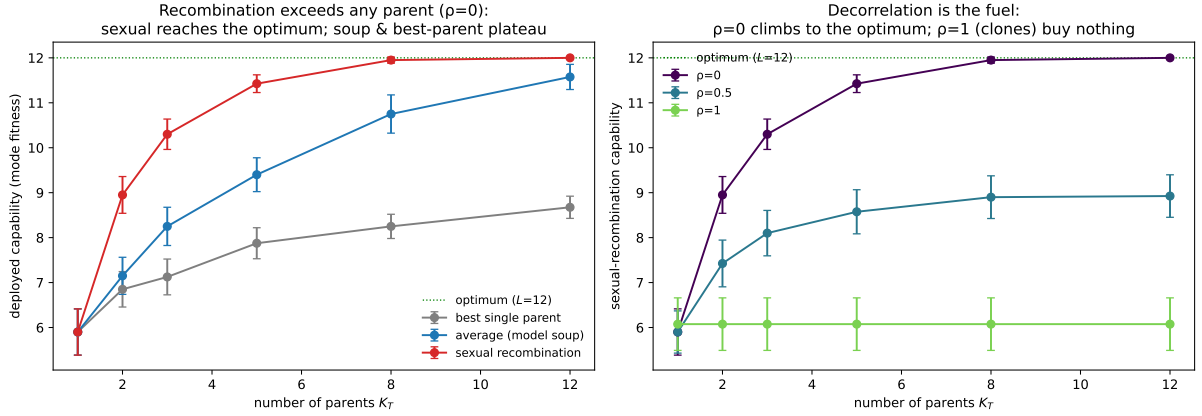
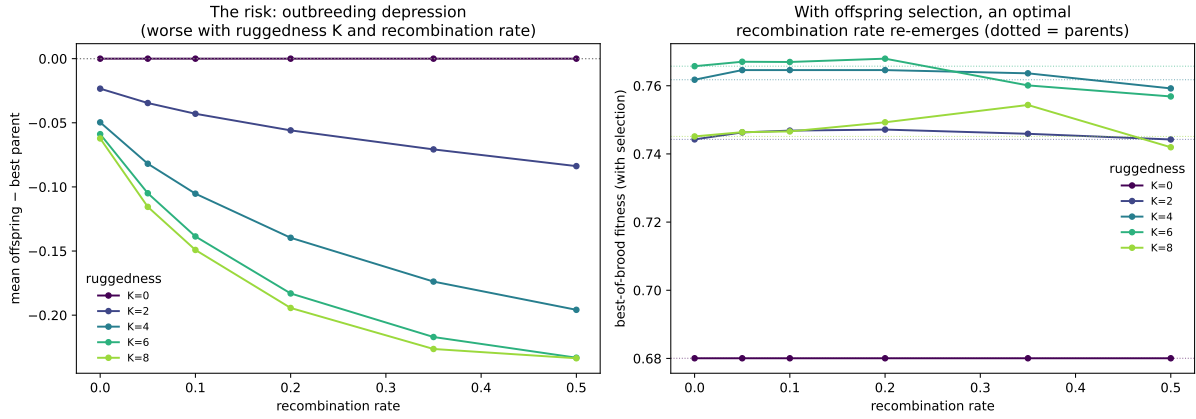
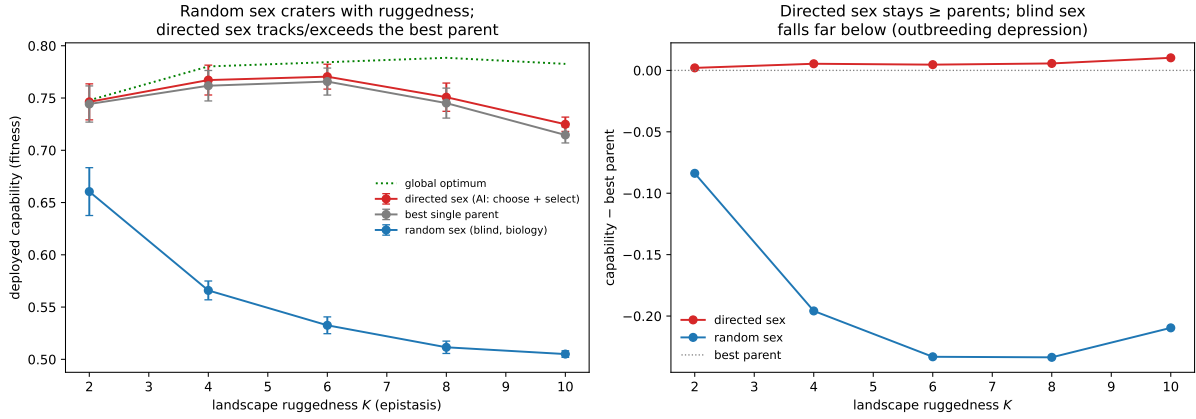


Figure 2: Recombination: the conservation law and the Fisher–Muller effect. (A, top) Refitting a child to the mean of its parents’ output distributions conserves rare-item mass at single-parent level regardless of parent count (blending inheritance); a strongest-source (union) operator realises the multi-parent gain. (B, bottom) Multi-locus recombination of decorrelated specialists assembles a genotype fitter than any parent, climbing to the optimum as parents are added, while the best single parent and the blended average plateau below.

E9 — landscape robustness: recombination helps when skills are complementary, but blindly merging entangled models causes outbreeding depression



E10 — directed sex beats biological sex: mate choice + offspring selection + unbounded parents rescue recombination where blind sex fails



E14 — monogamy vs promiscuity: the best mate-pool breadth shrinks as skills get more entangled

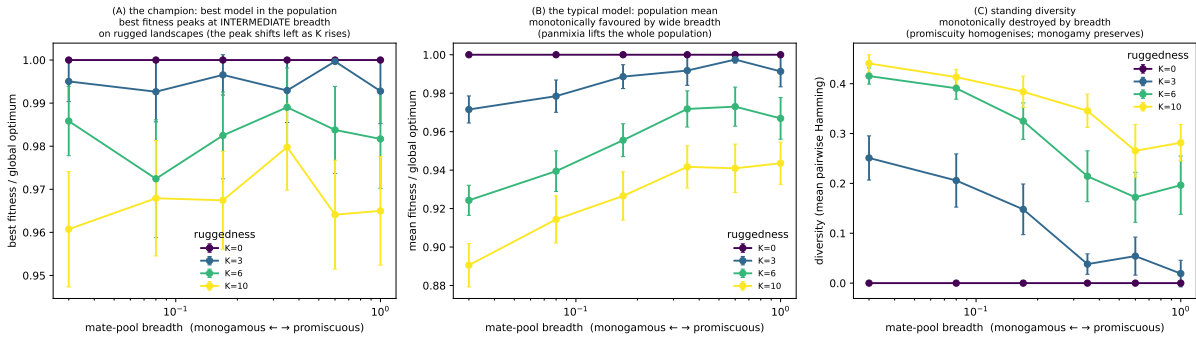


Figure 3: Rugged (epistatic) landscapes: risk, remedy, and structure. (A, top) Outbreeding depression: blind recombination of specialists drops offspring below their parents, worsening with ruggedness; the optimal recombination rate shrinks as skills entangle. (B, middle) Directed sex — unbounded parents, chosen mates, verifier-screened offspring — converts the catastrophe into a reliable gain at every ruggedness. (C, bottom) Mating structure: wide (promiscuous) mixing maximises the population mean but monotonically destroys diversity; the champion-optimal mate-pool breadth narrows as the landscape roughens.

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 — both couple the lineage to a non-drifting external signal — but they are different operators, and we 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) 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: changes harmless on their own background but deleterious in combination. A merged model is exactly the exposed hybrid. We 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 (17) — noting 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 (23), and richer symmetry groups remove more (24). We 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. 5B): two networks trained from different initialisations on the *same* task have a barrier that this alignment removes essentially entirely (residual ≈ 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 strongest constraint comes from the pre-registered **emergent test**: true BDM incompatibilities are emergent (each lineage’s changes harmless alone), so we let children diverge with *no conflicting signal anywhere* — 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 ≈ 0.50 , merge ≈ 0.955 — a sustained Fisher–Muller rescue). The same double result appears at the language-model tier (Fig. 5C): 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→12 epochs 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 we 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 (21), 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.

The supported conclusion, stated conditionally: **across this controlled grid, pre-merge functional disagreement predicted merge penalties (clustered bootstrap CIs excluding zero; held-out leave-one-condition-out $\rho \approx 0.35\text{--}0.40$), 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 ρ stable at $+0.37$ to $+0.53$ in each seed alone, geometry ≈ 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 ± 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→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.

Discussion

Design rules. Read as engineering, the results compress into rules an operator of a model population can apply. *Ground every generation* in verified reality — a few percent retained most diversity in our 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 our 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 we tested, what broke merging was conflicting conventions on shared circuitry, which is the thing to detect.

What this offers continual learning. Read into the field where these results most directly land: (i) a first-principles account of the **replay ratio** — the folklore constants ($\approx 1\%$, 5% , 25% ; 52, 53) acquire an equilibrium theory and a sharper prediction, that the required fraction is set by the rarest capability one refuses to lose (the $1 - e^{-m \cdot p}$ law), not by average loss — directly 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 (47, 48, 12, 62); (iii) **pre-merge interference prediction with a mechanism:** where the current state of the art fits regressions over candidate metrics (21), 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 we note the standing objection that apparent forgetting can be skewed task-inference over latent capability rather than erasure (64); our irreversibility results concern oracle-measured behavioural distributions, and distinguishing latent from extinct capability at language-model scale is an open and, we think, decisive experiment for both readings.

What is borrowed and what is ours. The diagnosis — collapse as drift — was published first by others and we cite it so (6–9), while noting the derivations are independent and convergent; prior art in the strict sense 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 (1, 19, 20), that model populations can climb (2–5), and that merge success admits ML-native predictors (21, 22). Ours 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. We claim the framework generated these measurements and experiments; we do not claim 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 we trust, and institutionalising that anchor (verification, replication, and challenge among models) is the society-level problem we pose but do 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.

Materials and Methods

Analytic tier. Pure NumPy/SciPy Wright–Fisher simulator over K-item distributions (knowledge as `p_t`; Zipf-tailed truth `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. Yadav P, Tam D, Choshen L, Raffel C, Bansal M (2023) TIES-Merging: resolving interference when merging models. *NeurIPS*. arXiv:2306.01708.
2. Akiba T, Shing M, Tang Y, Sun Q, Ha D (2025) Evolutionary optimization of model merging recipes. *Nat Mach Intell* 7:195–204.
3. GENOME: Nature-inspired population-based evolution of large language models (2025). arXiv:2503.01155.
4. Sakana AI (2025) Competition and attraction improve model fusion (M2N2). *GECCO*. arXiv:2508.16204.
5. Subramaniam V, Du Y, Tenenbaum JB, Torralba A, Li S, Mordatch I (2025) Multiagent finetuning: self-improvement with diverse reasoning chains. arXiv:2501.05707.
6. Shumailov I, et al. (2024) AI models collapse when trained on recursively generated data. *Nature* 631:755–759.
7. Riis S (2026) Drift and selection in LLM text ecosystems. arXiv:2604.08554.
8. Benati M, Londei A, Lanzieri D, Loreto V (2025) First-extinction law for resampling processes. arXiv:2509.20101.

9. Yoon Y, Hu D, Weissburg I, Qin Y, Jeong H (2025) Model collapse in the self-consuming chain of diffusion finetuning: a quantitative trait modeling perspective. *ICLR*. arXiv:2407.17493.
10. Muller HJ (1964) The relation of recombination to mutational advance. *Mutat Res* 1:2–9.
11. Gerstgrasser M, et al. (2024) Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. arXiv:2404.01413.
12. Yi B, Liu Q, Cheng Y, Xu H (2025) Escaping model collapse via synthetic data verification. arXiv:2510.16657.
13. Wright S (1931) Evolution in Mendelian populations. *Genetics* 16:97–159.
14. Fisher RA (1930) *The Genetical Theory of Natural Selection* (Clarendon, Oxford).
15. Muller HJ (1932) Some genetic aspects of sex. *Am Nat* 66:118–138.
16. Orr HA (1995) The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics* 139:1805–1813.
17. Orr HA, Turelli M (2001) The evolution of postzygotic isolation: accumulating Dobzhansky–Muller incompatibilities. *Evolution* 55:1085–1094.
18. Livnat A, Papadimitriou C (2016) Sex as an algorithm: the theory of evolution under the lens of computation. *Commun ACM* 59(11):84–93.
19. Yu L, Yu B, Yu H, Huang F, Li Y (2023) Language models are super Mario: absorbing abilities from homologous models (DARE). arXiv:2311.03099.
20. Wortsman M, et al. (2022) Model soups: averaging weights of multiple fine-tuned models. *ICML*. arXiv:2203.05482.
21. Zhou L, Zhao B, Yu R, Rodolà E (2026) Demystifying mergeability: interpretable properties to predict model merging success. arXiv:2601.22285.
22. Cao Y, Ran D, Guo Y, Wu M, Chen S, et al. (2026) An empirical study and theoretical explanation on task-level model-merging collapse. arXiv:2603.09463.
23. Ainsworth S, Hayase J, Srinivasa S (2022) Git Re-Basin: merging models modulo permutation symmetries. arXiv:2209.04836.
24. Li T, Shen Z (2026) Scaling linear mode connectivity and merging to billion-parameter pretrained transformers. arXiv:2606.23607.
25. Kauffman SA, Levin S (1987) Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* 128:11–45.
26. Lehman J, Stanley KO (2011) Abandoning objectives: evolution through the search for novelty alone. *Evol Comput* 19:189–223.
27. Pari J, Jelassi S, Agrawal P (2024) Collective model intelligence requires compatible specialization. arXiv:2411.02207.
28. Hu EJ, et al. (2021) LoRA: low-rank adaptation of large language models. arXiv:2106.09685.
29. Sharma E, Roy DM, Dziugaite GK (2024) The non-local model merging problem: permutation symmetries and variance collapse. arXiv:2410.12766.

30. Kozodoi N, Afolabi Z, Butler J (2026) Are we merging the right models? Impact of expert training duration on model merging for LLMs. *arXiv:2607.11997*.
31. Laufer B, Oderinwale H, Kleinberg J (2025) Anatomy of a machine learning ecosystem: 2 million models on Hugging Face. *arXiv:2508.06811*.
32. Horwitz E, Shul A, Hoshen Y (2025) Unsupervised model tree heritage recovery. *ICLR*. *arXiv:2405.18432*.
33. Jiang W, et al. (2024) PeaTMOSS: a dataset and initial analysis of pre-trained models in open-source software. *MSR*. *arXiv:2402.00699*.
34. Villalobos P, Ho A, Sevilla J, Besiroglu T, Heim L, Hobbhahn M (2024) Position: will we run out of data? Limits of LLM scaling based on human-generated data. *ICML*. *arXiv:2211.04325*.
35. Thompson B, et al. (2024) A shocking amount of the web is machine translated. *Findings of ACL*. *arXiv:2401.05749*.
36. Liang W, et al. (2024) Monitoring AI-modified content at scale. *ICML*. *arXiv:2403.07183*.
37. Goddard C, et al. (2024) Arcee’s MergeKit: a toolkit for merging large language models. *EMNLP Industry Track*, 477–485. *arXiv:2403.13257*.
38. Yang E, et al. (2024) Model merging in LLMs, MLLMs, and beyond: methods, theories, applications and opportunities. *arXiv:2408.07666*.
39. Brinkmann L, et al. (2023) Machine culture. *Nat Hum Behav* 7:1855–1868.
40. Park JS, O’Brien JC, Cai CJ, et al. (2023) Generative agents: interactive simulacra of human behavior. *UIST*. *arXiv:2304.03442*.
41. Guo T, Chen X, Wang Y, et al. (2024) Large language model based multi-agents: a survey of progress and challenges. *IJCAI*. *arXiv:2402.01680*.
42. Tomasev N, Franklin M, Leibo JZ, et al. (2025) Virtual agent economies. *arXiv:2509.10147*.
43. Adler B, et al. (2024) Nemotron-4 340B technical report. *arXiv:2406.11704*.
44. Abdin M, et al. (2024) Phi-4 technical report. *arXiv:2412.08905*.
45. McCloskey M, Cohen NJ (1989) Catastrophic interference in connectionist networks. *Psychol Learn Motiv* 24:109–165.
46. French RM (1999) Catastrophic forgetting in connectionist networks. *Trends Cogn Sci* 3:128–135.
47. Robins A (1995) Catastrophic forgetting, rehearsal and pseudorehearsal. *Connect Sci* 7:123–146.
48. Shin H, Lee JK, Kim J, Kim J (2017) Continual learning with deep generative replay. *NeurIPS*. *arXiv:1705.08690*.
49. McClelland JL, McNaughton BL, O’Reilly RC (1995) Why there are complementary learning systems in the hippocampus and neocortex. *Psychol Rev* 102:419–457.
50. Kumaran D, Hassabis D, McClelland JL (2016) What learning systems do intelligent agents need? *Trends Cogn Sci* 20:512–534.

51. Schwarz J, et al. (2018) Progress & Compress: a scalable framework for continual learning. *ICML*.
52. Ibrahim A, et al. (2024) Simple and scalable strategies to continually pre-train large language models. *TMLR*. arXiv:2403.08763.
53. Scialom T, Chakrabarty T, Muresan S (2022) Fine-tuned language models are continual learners. *EMNLP*. arXiv:2205.12393.
54. Biderman D, et al. (2024) LoRA learns less and forgets less. *TMLR*. arXiv:2405.09673.
55. Ilharco G, et al. (2023) Editing models with task arithmetic. *ICLR*. arXiv:2212.04089.
56. Marczak D, et al. (2024) MagMax: leveraging model merging for seamless continual learning. *ECCV*. arXiv:2407.06322.
57. Alexandrov A, et al. (2024) Mitigating catastrophic forgetting in language transfer via model merging. *Findings of EMNLP*. arXiv:2407.08699.
58. Dziadzio S, et al. (2025) How to merge your multimodal models over time? *CVPR*. arXiv:2412.06712.
59. Toneva M, et al. (2019) An empirical study of example forgetting during deep neural network learning. *ICLR*. arXiv:1812.05159.
60. Kandpal N, et al. (2023) Large language models struggle to learn long-tail knowledge. *ICML*.
61. Liu X, et al. (2022) Long-tailed class incremental learning. *ECCV*. arXiv:2210.00266.
62. Feng Y, et al. (2024) Beyond model collapse: scaling up with synthesized data requires verification. arXiv:2406.07515.
63. Crutchfield JP, Whalen S (2012) Structural drift: the population dynamics of sequential learning. *PLoS Comput Biol* 8:e1002510.
64. Kotha S, Springer JM, Raghunathan A (2024) Understanding catastrophic forgetting in language models via implicit inference. *ICLR*.
65. Rusu AA, et al. (2016) Progressive neural networks. arXiv:1606.04671.

E11 — the dynamic Lamarckian society: grounding + directed sex + diversity climb to the optimum; remove any one and it breaks (the vertical claim, C3)

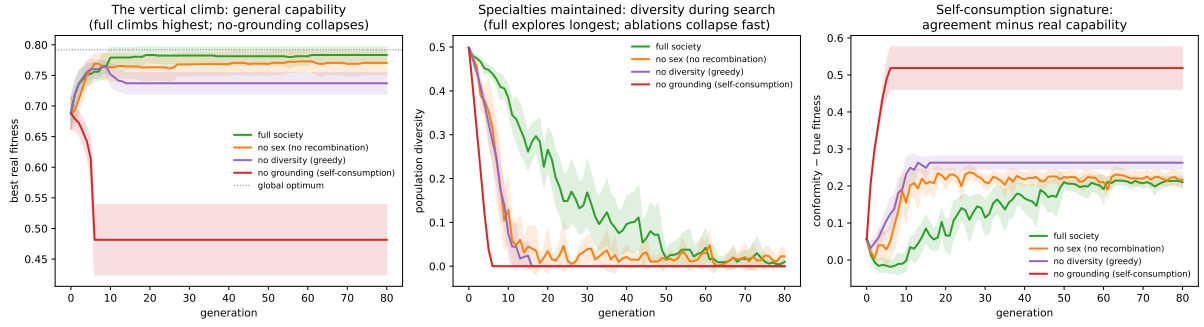
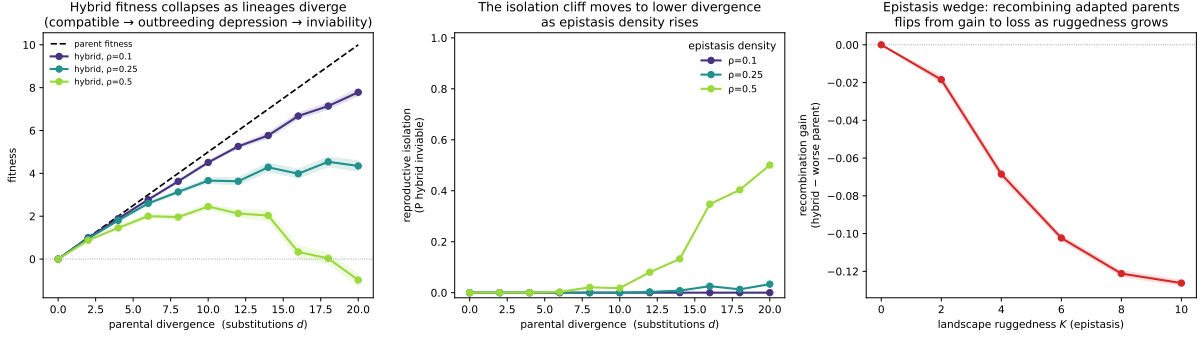
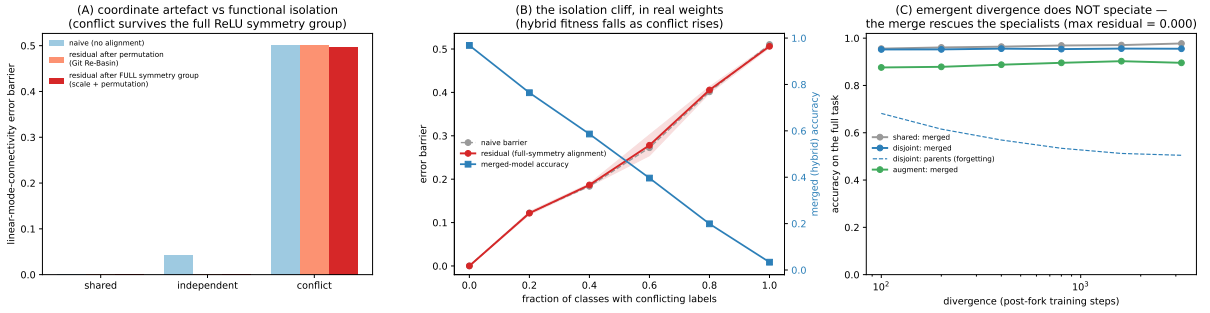


Figure 4: The society: grounding, sex, and diversity are jointly necessary. A finite agent population on a rugged NK landscape under a grounded selection score. Four-arm ablation: the full system climbs to near the global optimum; removing grounding collapses the population onto a confident, unfit consensus (self-consumption); removing recombination strands it on local optima; removing diversity converges it prematurely. Each ablation fails differently.

E12 — model speciation: when two diverged models are too incompatible to merge



E13 — real-weight model speciation: isolation requires functional conflict; alignment (even modulo the full symmetry group) cannot remove it, and compatible specialists merge into a rescuing generalist



LLM-tier model speciation: conflict provokes function-specific hybrid breakdown; no isolation emerges from duration alone (shared base controls coordinate mismatch — a cleaner test of conflict-associated failure)

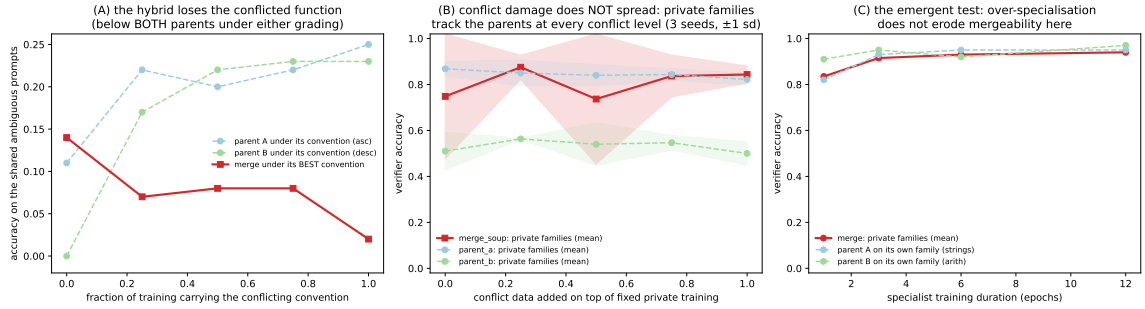


Figure 5: Model speciation across three tiers. (A, top) Analytic: hybrid fitness traces compatible → outbreeding depression → inviability; the cliff arrives earlier the denser the incompatibilities; incompatibility count snowballs with divergence. (B, middle) Trained MLPs: the merge barrier decomposed under the complete unit symmetry group — same-task/different-init barriers are coordinate artefacts (removed by alignment); conflicting-task barriers survive in full, with hybrid fitness falling $0.97 \rightarrow 0.03$; divergence without conflict produced no isolation, the merge instead rescuing the forgetting specialists. (C, bottom) Language models: conflicting conventions produce function-specific hybrid breakdown; over-training disjoint specialists produces none — at every tier tested, isolation had to be provoked by functional conflict.

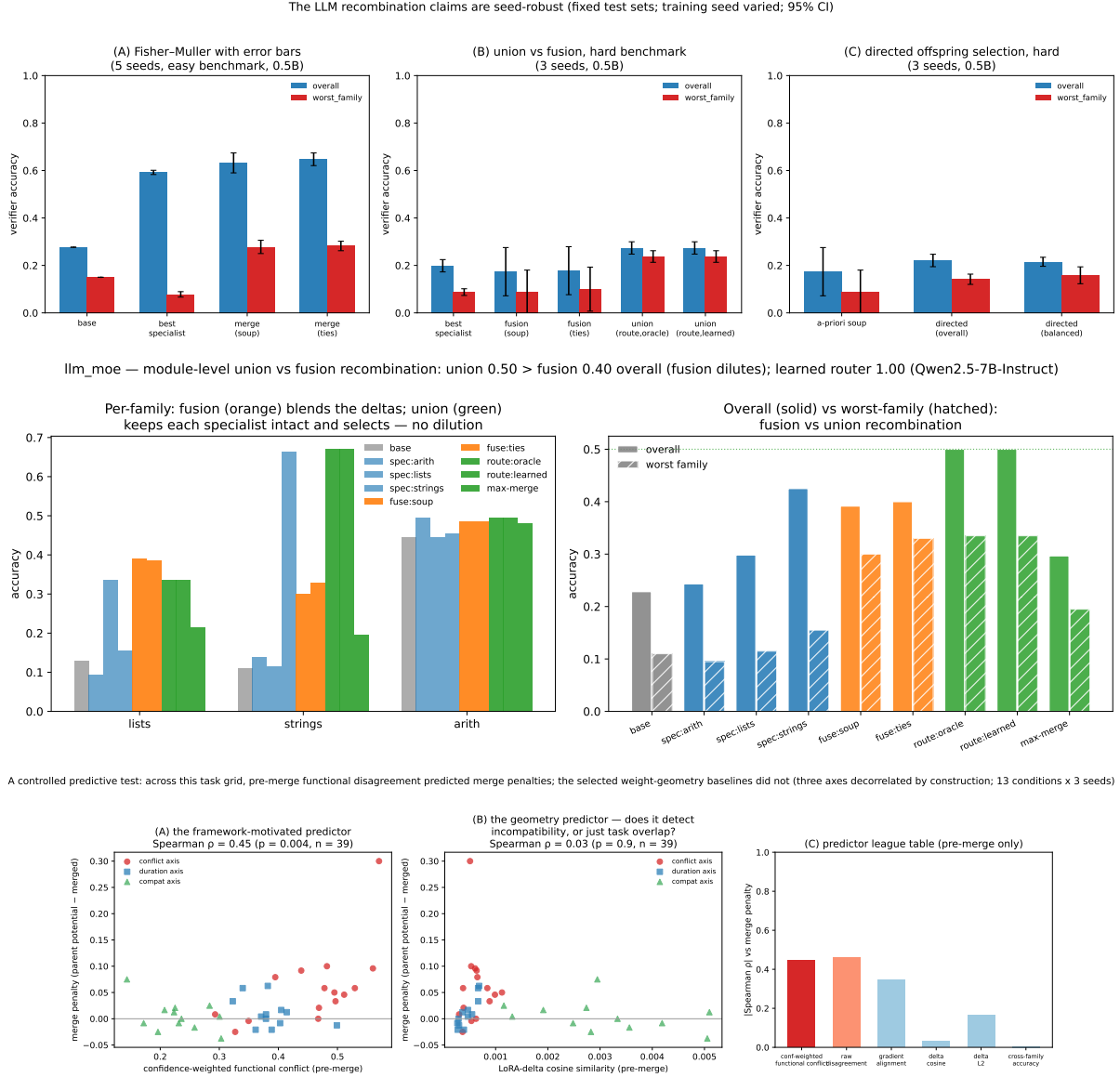


Figure 6: The language-model tier. (A, top) Seed-replicated recombination claims (fixed test sets, training seed varied, 95% CI): merges beat every specialist; union-preserving routing and directed offspring selection beat the blend in every seed on headroom tasks, including one catastrophic blend failure they avoided. (B, middle) The headroom rule at 7B on hard (unsaturated) tasks: the weight-average dilutes a fragile specialist below the best single parent; routing preserves it. (C, bottom) The controlled predictive test: across a task grid with conflict, compatible-overlap, and duration axes decorrelated by construction, pre-merge functional disagreement predicts merge penalty (held-out $\rho \approx 0.4$) while weight-geometry baselines show no detectable association; paired predictor differences are not individually significant.