

The evolution of sex for artificial intelligence: the figures explained

Appendix to the paper. Written for a reader with A-level biology or maths and no background in machine learning.

Why this appendix exists

The paper sits between two fields. Its questions and its theory come from population genetics; its experiments are machine learning, run on simulations, small neural networks and language models. A biologist can follow the argument in the main text while still finding the experimental details opaque: what a model is trained on, what an adapter is, what a verifier measures, why a result rests on seeds rather than replicates. This appendix answers those questions figure by figure. It repeats every figure of the paper, main and supplementary, with a legend that explains the experiment behind it in plain terms, so that a reader from biology can judge the evidence and not only the analogy.

How to read this document

The paper asks one question: when artificial-intelligence models are built from other models, what happens to what they know over the generations? Modern AI systems are rarely trained from nothing. A new model is usually a copy of an older one that has been trained a little further (a *child* of a *parent*), it is often trained on text that earlier models wrote, and two trained models are often blended into one by averaging the numbers inside them (*merging*). Those three habits give AI models parents, siblings and descendants, and biology has a hundred years of theory about populations like that: *population genetics*, the mathematics of how genes spread, vanish and recombine over generations.

The paper takes that theory literally. It treats a model's knowledge as a set of *items* (a fact, a skill, a habit of answering), each with a frequency, exactly as a population geneticist treats *alleles* (the alternative versions of a gene) and their frequencies. It then tests, at three levels of realism, whether the biological rules hold. The three levels are: a pure simulation with known answers (called the *inheritance model*), small neural networks trained on data the authors fully control, and real language models (the family of systems behind chatbots). Every figure below belongs to one or more of those levels.

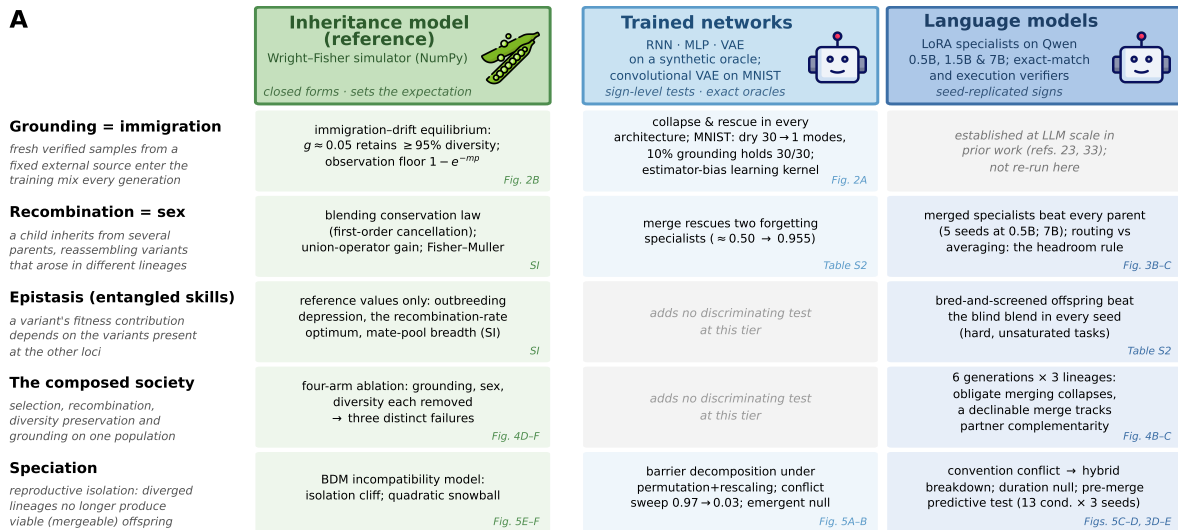
A few terms come back in every figure:

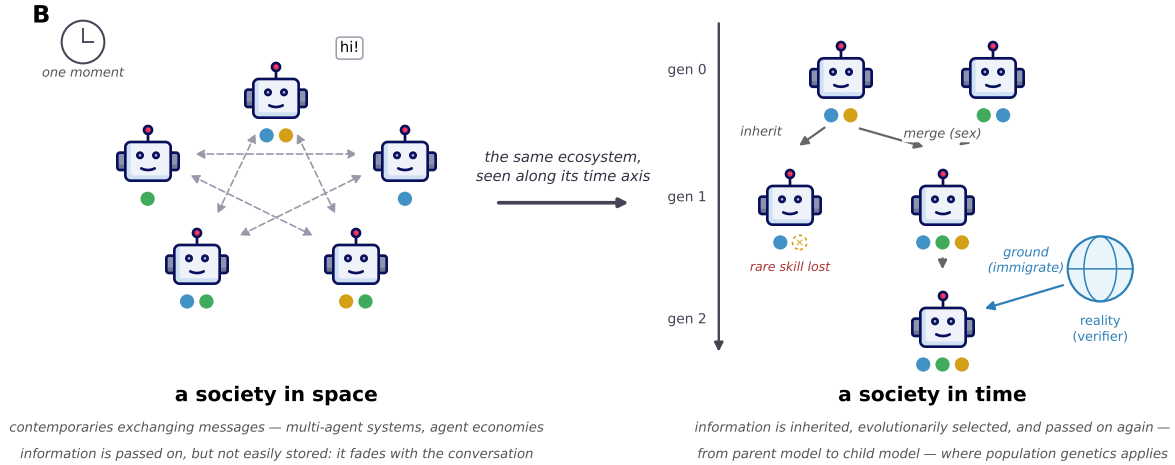
- **Generation.** One round of “train a child from its parent”. A lineage that goes through ten rounds has ten generations.
- **Grounding.** Mixing some genuine, checked real data into what a child is trained on, instead of training it only on what its parent produced. In the biological reading this is *immigration*: new individuals arriving from outside.
- **Model collapse.** The gradual loss of rare knowledge when each generation is trained only on the previous one. In biology the same process is called *genetic drift*: in a small population, rare alleles disappear by chance, not because anything selects against them, in the way that rare surnames die out in a small village.

- **Diversity, or heterozygosity, H.** A number between 0 and 1 that measures how spread out the knowledge is. If you pick two items at random, H is the chance they differ. H near 1 means many items share the frequency; H equal to 0 means one item has taken over.
- **Merging.** Making a new model by averaging the internal numbers (the *weights*) of two or more trained parents. The paper's central claim is that this is the AI counterpart of *sexual reproduction*, with the same benefits and the same dangers.
- **Verifier.** A program that can mark an answer right or wrong automatically (run the code, check the arithmetic, compare with the known answer). It is the paper's stand-in for "reality that can say no".
- **Accuracy.** The fraction of test questions a model gets right, from 0 to 1.
- **Seed.** Training a neural network involves random choices. Repeating an experiment with a different random seed and getting the same answer shows the result is not a fluke. Error bars in the figures are 95% confidence intervals over seeds or replicates.

Each legend below says what was done, what you are looking at, and what it means.

Figure 1. A map of the whole study



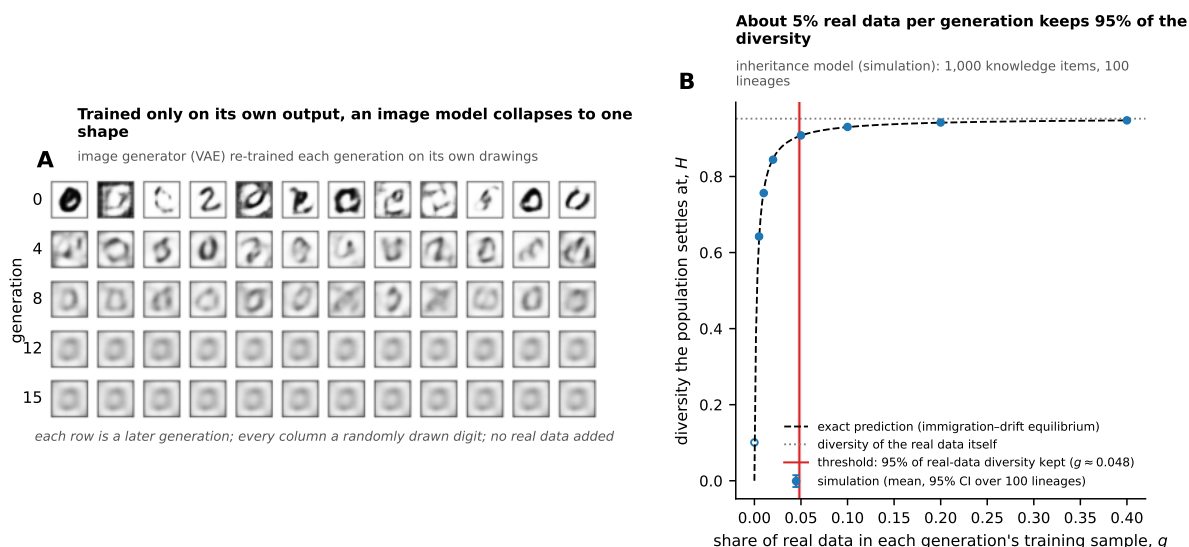


What was done. Nothing is measured in this figure. It is a map. Panel A is a grid: each row is one biological mechanism the paper borrows, each column is one of the three levels of realism at which it was tested. The rows are grounding (immigration), recombination (sex), epistasis (skills that only work in combination), the complete “society” with all mechanisms running at once, and speciation (when two lineages can no longer produce a working hybrid). The columns are the inheritance model (a simulation with exact answers, in green), trained small networks (blue), and language models (blue). Each filled cell names what was run there and, in the corner, which figure reports it. Cells marked “no counterpart” or “established in prior work” were deliberately not run: the paper tests each claim at the cheapest level that could prove it wrong, and moves to a more expensive level only when that adds a new test rather than a repeat.

Panel B shows the change of viewpoint the whole paper rests on. On the left is how people usually picture a group of AI models: contemporaries exchanging messages, a *society in space*. On the right the same group is drawn along its time axis: a model inherits from a parent, merges with a partner, receives fresh real data, and passes the result on. That is a *society in time*, and it is exactly the kind of object population genetics was built to describe. The coloured dots on each robot are its skills. Follow the gold dot: it is rare, it is lost when a child inherits from a single parent, it is recovered when two complementary parents merge, and it is re-supplied by grounding (the globe).

What it means. If you remember one thing from this figure, remember the right-hand side of panel B. The rest of the paper is a list of what happens to the gold dot.

Figure 2. How much real data stops collapse

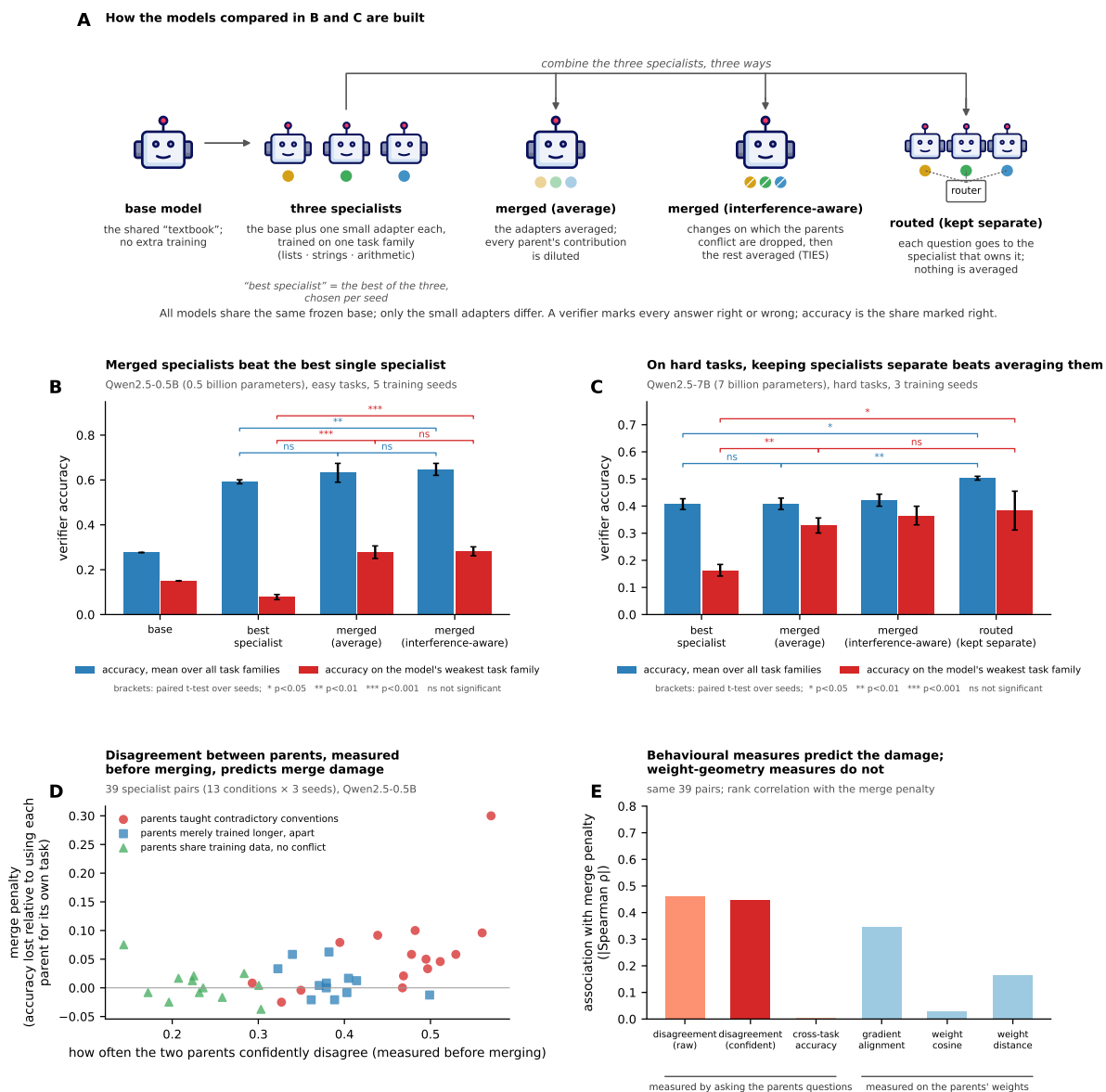


What was done. Two experiments on the same question. In panel A a small image-generating network (a *variational autoencoder*, a type of network that learns to draw new examples of what it was shown) was trained on handwritten digits, then a fresh copy was trained only on the digits the first one drew, then another on that one's output, for fifteen generations. The digits were sorted into thirty kinds (each digit in three stroke thicknesses), some kinds common and some rare, and an independent classifier checked which kind each drawn digit belonged to. In panel B the same question was asked of the inheritance model, the pure simulation: a population of 1,000 knowledge items, 200 samples drawn per generation, and a fraction g of fresh real samples mixed in each time, swept from 0 to 0.4 across 100 independent lineages.

What you see. In A, each row is a generation (0, 4, 8, 12, 15) and each column a randomly chosen drawing. With no real data added, the drawings degrade from recognisable digits to one blurred grey shape: the population has collapsed to a single kind. In B, the vertical axis is the diversity H the population settles at, and the horizontal axis is the grounding fraction g . Points are the simulation, the dashed line is the exact mathematical prediction, and the dotted line is the diversity of the real data itself. The red line marks the g at which the population keeps 95% of the real data's diversity for ever: about 0.05, one sample in twenty.

What it means. A small, steady trickle of checked real data is enough to hold the population's diversity indefinitely, much as a few migrants per generation keep an island population healthy. The curve is smooth: there is no sudden switch between "safe" and "collapsing", so any threshold you quote is a choice of how much diversity you want to keep. The image network needed about twice the simulation's fraction (10% rather than 5%) because a trained network is not the ideal copier the simulation assumes. The text also explains why the *rarest* items need more than this: to have a fair chance of seeing an item that occurs once in ten thousand real examples, you need about ten thousand real examples every generation.

Figure 3. Merging language models: when it helps, and predicting when it will hurt



What was done. Every data panel uses small language models trained to be *specialists*: starting from one shared base model (Qwen2.5, of 0.5 or 7 billion parameters), a small add-on set of weights called a *LoRA adapter* is trained on one family of tasks (list puzzles, string puzzles or arithmetic). Think of the base model as a shared textbook and each adapter as one student's margin notes. Merging two specialists means averaging their notes. A verifier marks every answer.

Panel A is a picture of the five kinds of model the next two panels compare, left to right: the base alone; the three specialists (the "best specialist" is whichever of the three scores highest, chosen separately for each seed); the three combined by averaging their adapters, which dilutes each parent's contribution; the three combined after first dropping the changes on which the parents disagree (an "interference-aware" merge); and routing, which keeps the three specialists intact and sends each question to the one that owns it.

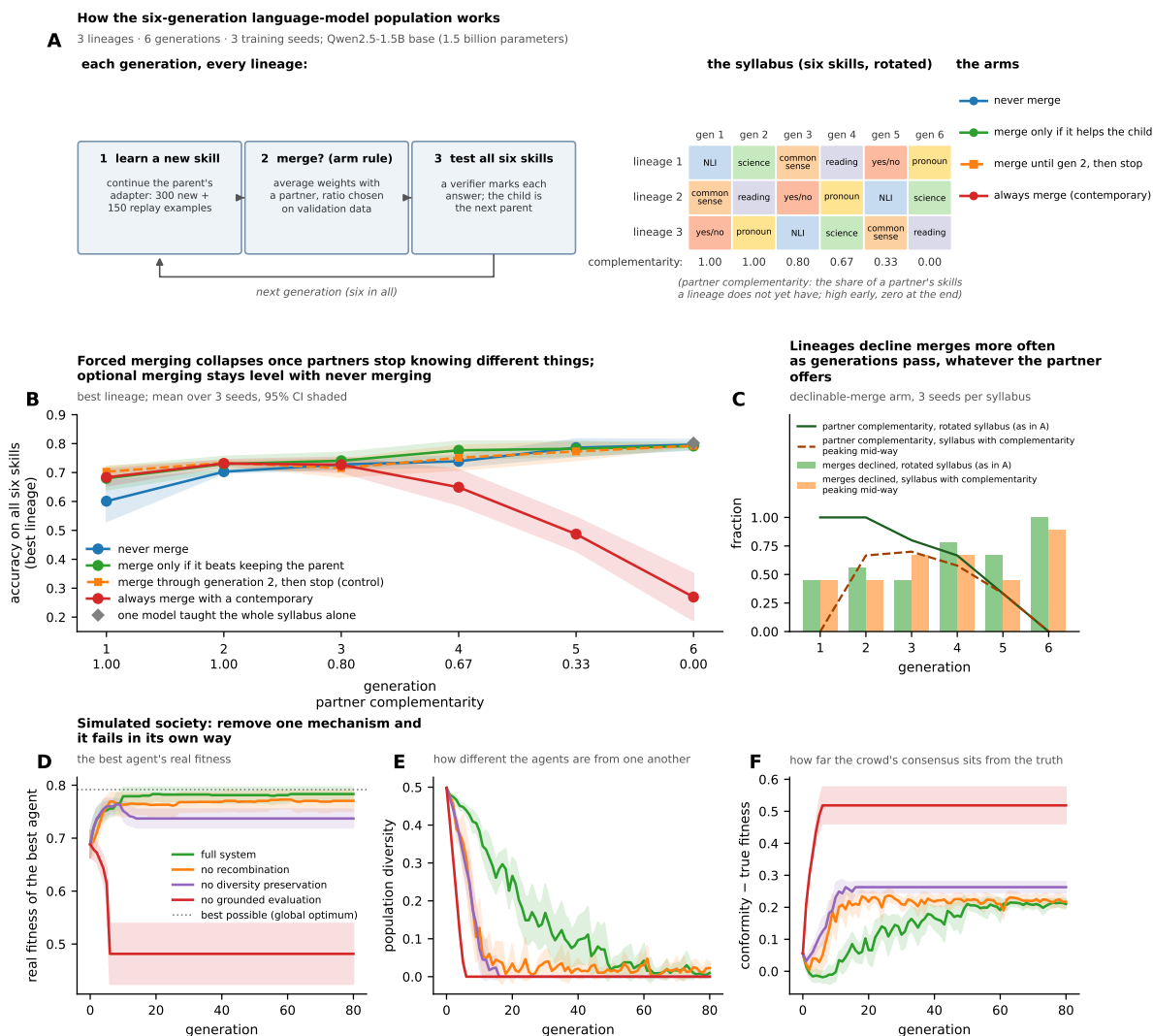
Panels B and C. Bars show accuracy averaged over all task families (blue) and on the family

each model is worst at (red); brackets mark pairs of bars that differ significantly across seeds (stars) or do not (ns). In B (0.5B model, five seeds) the merged models beat the best single specialist overall, and they are the only models that are competent on every family at once. This is the AI version of what geneticists call the *Fisher–Muller effect*: sex gathers into one offspring useful variants that arose in different individuals. In C the tasks were made deliberately hard so that a larger 7B model was not already at the ceiling. Here plain averaging only matches the best specialist, because averaging dilutes each specialist’s own skill, while *routing* (keeping the specialists separate and sending each question to the right one) wins by a wide margin. The rule the paper draws is about *headroom*: keep specialists separate whenever the average falls short of what they could jointly do.

Panels D and E. Can you tell in advance whether a merge will go badly? Thirty-nine pairs of specialists were built along three axes: pairs trained to answer the same questions in *contradictory* ways (red), pairs trained on the same questions in the *same* way (green), and pairs simply trained for longer on different things (blue). Before merging, six quantities were measured on each pair. The one on the horizontal axis of D is functional conflict: how often the two parents confidently disagree when asked the same probe questions. The vertical axis is the *merge penalty*: how far the merged model falls short of what the pair could have scored if each question went to the parent that knew it. The penalty concentrates in the red points. Panel E compares the six predictors: how well each one ranks the pairs by penalty. Disagreement measured by asking questions predicts damage; measures of how far apart the parents’ internal numbers are (weight cosine, weight distance) do not.

What it means. Merging complementary specialists can produce a model better than any of its parents, and a cheap behavioural test on the parents forecasts when merging will fail. The green points carry a warning for other researchers: pairs that share training data have similar weights and also merge worse, so a predictor based on weight similarity can look good for the wrong reason.

Figure 4. A population of language models over six generations



What was done. Panel A is a picture of the set-up; panels B and C follow three lineages of language models for six generations. Each generation, every lineage learns one new skill from a public dataset (six in total: reasoning about sentences, science questions, common-sense completion, reading comprehension, yes/no questions, pronoun resolution) by continuing to train its parent's adapter, so what the parent learned passes on. The three lineages take the six skills in rotated orders, like three students working through one syllabus in different sequences, so early on a partner knows things you lack and late on it knows nothing you lack. That quantity, the share of a partner's skills you do not have, is called *complementarity* and is printed under the generation numbers. Between generations a lineage may merge with a partner. The arms differ in the rule: never merge; always merge with a contemporary; merge only if the merged model beats keeping the parent (a *declinable* merge); merge for the first three generations and then stop. Everything was repeated with three seeds.

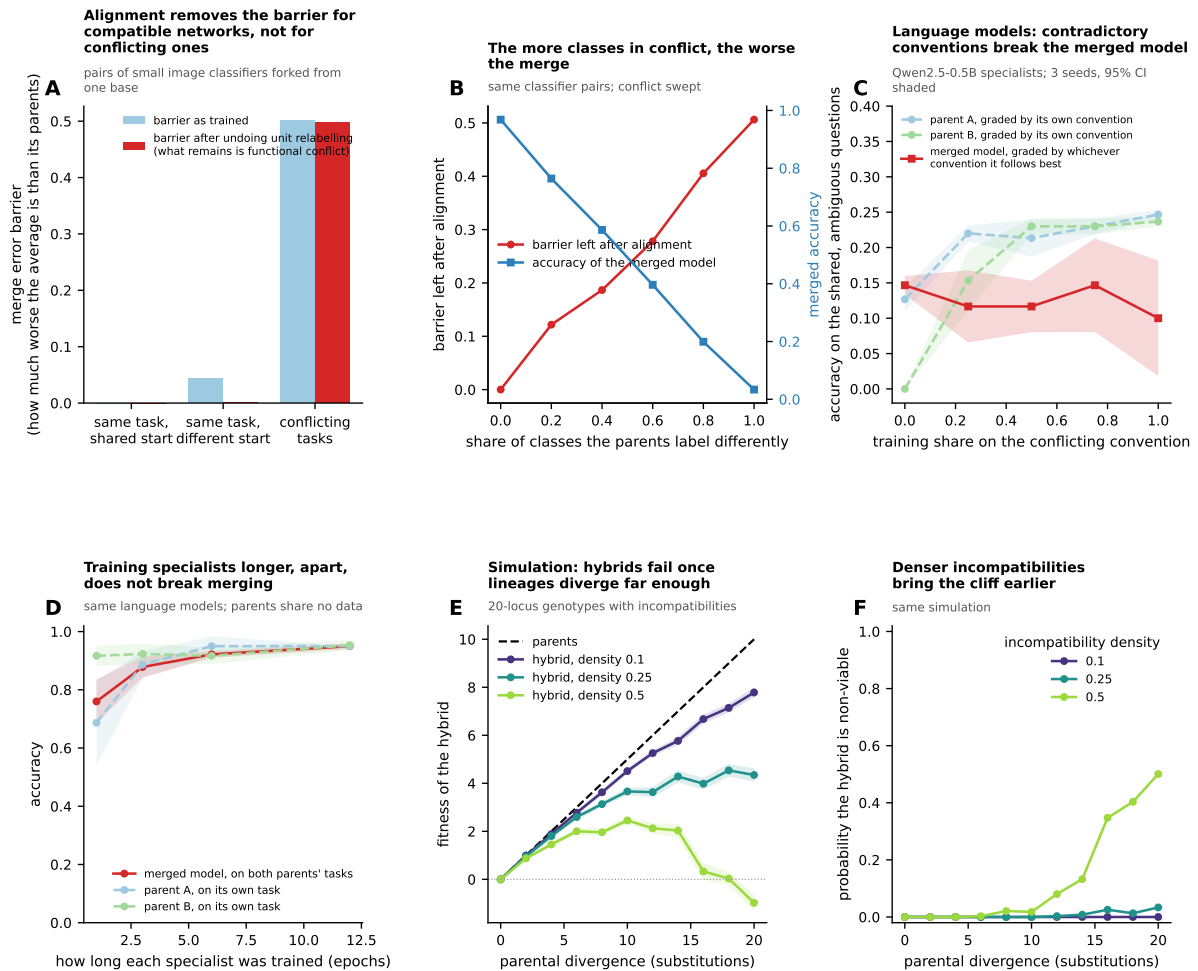
Panel A shows one generation as a loop (learn a new skill, decide whether to merge, take the test), the syllabus as a grid of three lineages by six generations with the skills colour-coded so the rotation is visible, and the four merging rules with the colours used in panel B. **Panel B.** Accuracy of the best lineage on all six skills, generation by generation. Never merging (blue) and the declinable merge (green) end level, near 0.80. Always merging (red) tracks them for three generations and then collapses to under 0.30, beginning when partners stop being

complementary. The orange dashed line is the “merge early, then stop” control, and the grey diamond is a single model taught the whole syllabus alone. **Panel C.** How often the declinable lineages refused a merge (bars) against complementarity (lines), under the rotated syllabus (green) and a second syllabus in which complementarity starts at zero, peaks in the middle and returns to zero (orange). Refusals rise with generation under both, and once generation is accounted for they do not follow complementarity.

Panels D to F are the simulation that motivated the design: sixty simulated agents evolving on a rugged fitness landscape (a landscape where a variant’s value depends on which other variants it sits next to), with all four mechanisms running (grounding, recombination, diversity preservation, mutation) and one removed per arm. Removing grounding (red) makes the population agree confidently on a wrong answer: it optimises fitting the crowd instead of reality. Removing recombination (orange) or diversity (purple) strands it lower. Each removal fails in its own way.

What it means. Merging with a partner that knows conflicting things is what destroys a population; giving each lineage the right to refuse a merge, or simply stopping early, avoids the collapse at no cost. The simulation shows why all the mechanisms are needed at once. Two things the paper had hoped to see were not seen: refusals did not track complementarity, and (Figure S15) adding survival of the fittest did not make merging lineages finish ahead.

Figure 5. Model speciation: when two lineages can no longer merge



What was done. In biology, two lineages pushed far enough apart become separate species: their hybrids fail, as a mule is sterile, because two genomes that each work cannot run together in one cell. The paper asks whether the same happens to models. The difficulty is a known nuisance: two networks trained separately can differ in their weights for a trivial reason. The internal units of a network can be renumbered, and scaled up and down in matching pairs, without changing what the network computes, so two networks that do the same job can look very different inside. *Alignment* undoes this relabelling before merging. Panels A and B use small networks whose units can be aligned exactly; panels C and D use language models; panels E and F use the simulation.

Panel A. The height of the bar is the *barrier*: how much worse the average of two networks is than the networks themselves. Two copies trained from different random starts on the same task have a barrier that alignment removes almost entirely (from 0.04 to about 0.001). Two networks trained on *conflicting* labels (the same images, some classes deliberately relabelled) have a barrier alignment leaves untouched (0.50), and the merged model is useless. **Panel B.** Sweeping the fraction of classes in conflict moves the merged model’s accuracy from 0.97 to 0.03: a cliff. **Panel C.** Language models: two specialists were given a shared set of ambiguous questions (“sort this list”, direction unstated) and taught opposite conventions (one sorts ascending, the other descending). As the share of such conflicting training grows, each parent stays good under its own convention, but the merged model’s accuracy under its best convention falls below both

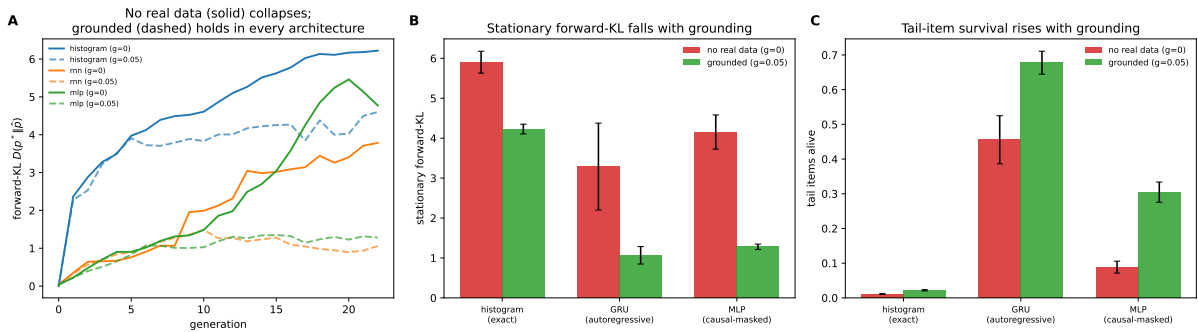
parents, in all three seeds (shaded bands). **Panel D.** The control: specialists trained for longer and longer on *different* tasks, with no conflict at all. The merged model gets better, never worse, however long the parents train. **Panels E and F.** The simulation: hybrid fitness tracks the parents while lineages are compatible and then crashes, sooner when incompatibilities are denser, and the chance of a non-viable hybrid rises with divergence.

What it means. What breaks merging is conflicting conventions on shared machinery, not distance or specialisation as such. Left alone, specialisation did not produce “species” in any experiment here; isolation had to be provoked by conflict. That is good news for anyone merging models, and it is why the pre-merge test in Figure 3D works.

Supplementary figures

The supplementary figures are the experiments behind the main text that either reproduce a known result, calibrate a method, or replicate a main result on more seeds or a second system. They keep the working titles the experiments were run under.

Figure S1. Collapse and rescue in three different kinds of network

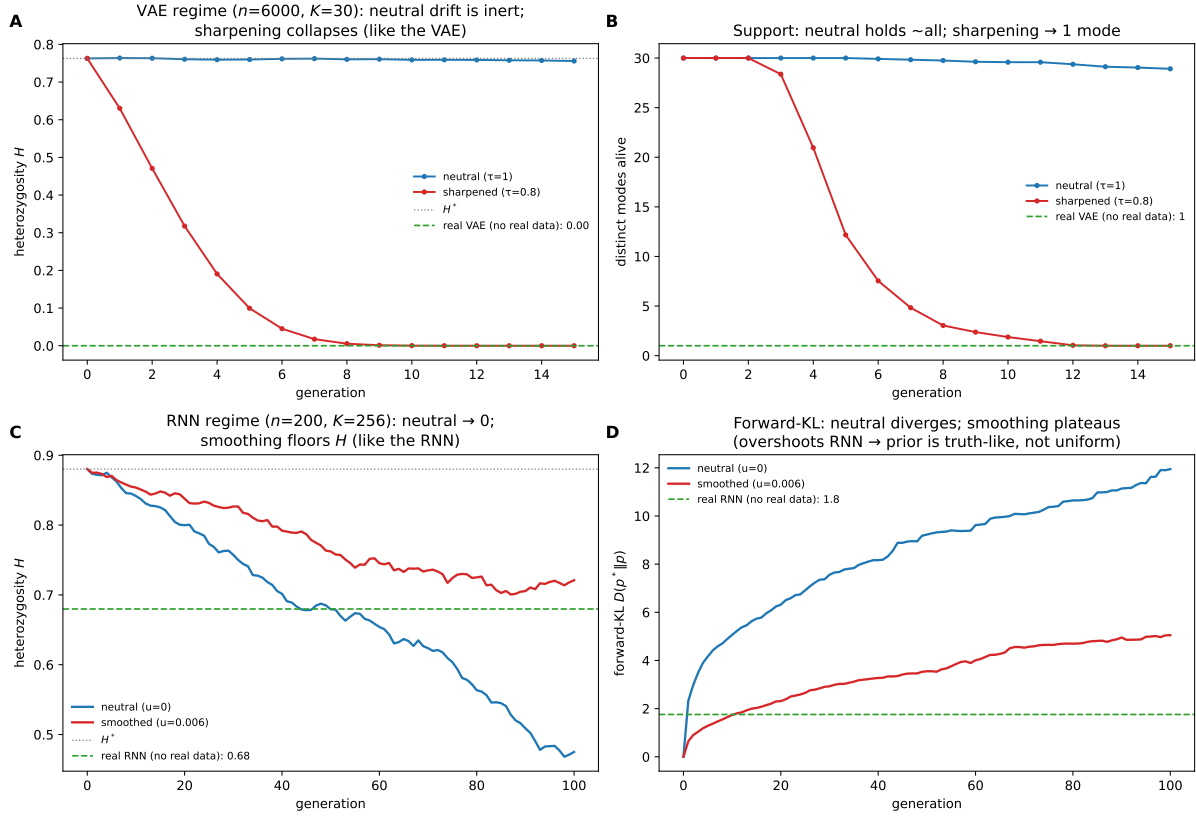


What was done. The same generational loop as Figure 2 (train a child only on its parent’s output, with or without 5% real data) was run with three generators: an exact histogram (a simple frequency count, no neural network), a recurrent neural network (one that reads and writes sequences one token at a time), and a feed-forward network. Each had to learn a synthetic “universe” of 256 kinds of item whose true frequencies were known exactly, for 22 generations, five times over.

What you see. (A) distance from the true distribution (a quantity called forward KL divergence, which grows the more of the truth a model fails to cover) against generation. Solid lines, with no real data, climb in every architecture; dashed lines, with 5% real data, stay low. (B, C) the same at the end of the run, as bars, and the fraction of rare items still alive.

What it means. Collapse and its rescue by grounding are not a quirk of one type of network. The histogram’s bars for rare items are tiny because a frequency count drops a rare item outright once it is unseen, whereas the neural networks keep some alive by “smoothing”, spreading a little probability onto things they have not seen. That difference is the subject of Figure S2.

Figure S2. Why real networks deviate from the ideal, in opposite directions

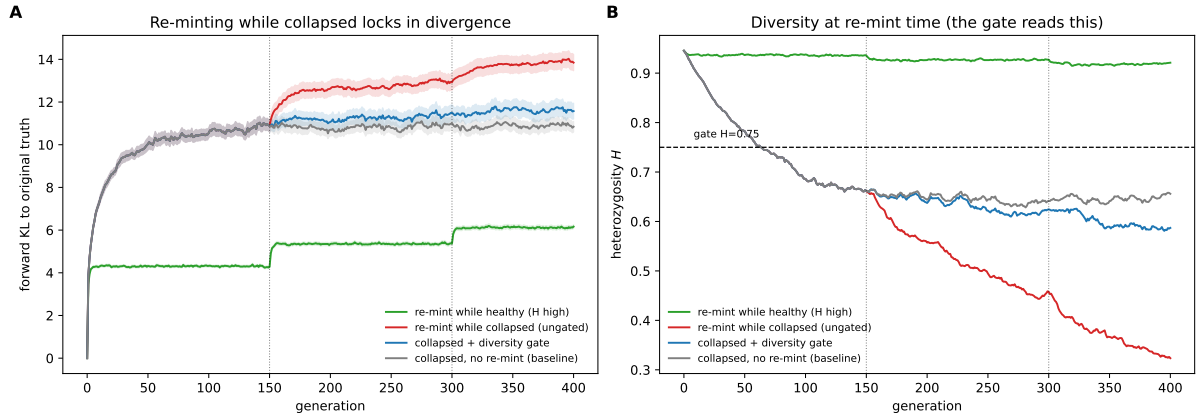


What was done. The inheritance model assumes a perfect copier: a child’s frequencies are exactly the frequencies it sampled from its parent. Real networks are not perfect copiers. This figure adds two knobs to the simulation’s copying step: a *smoothing* knob (a small pull toward treating all items as possible) and a *sharpening* knob (a temperature that concentrates probability on the commonest items), and asks whether either reproduces what the real networks did.

What you see. Blue is the ideal copier, red the copier with one knob turned, green dashed the level the real trained network actually reached. Panels A and B: the image network of Figure 2 (6,000 samples per generation, 30 kinds). The ideal copier barely drifts at that sample size, yet the real network collapsed to a single kind; turning the sharpening knob reproduces the collapse. Panels C and D: the recurrent network (200 samples, 256 kinds). The ideal copier drives diversity to zero, yet the real network keeps a floor of diversity; turning the smoothing knob reproduces the floor.

What it means. A trained network behaves like the textbook model of drift plus a bias that depends on its architecture: some networks add collapse, some resist it. In biological terms the two knobs are different things. The smoothing knob is recurrent mutation: variants appear in the child that it did not inherit, though here they are the network’s own inventions rather than real knowledge, which is why counting them overstates its health. The sharpening knob is not mutation at all; it is selection in favour of whatever is already common, which removes variants and never creates them. Knowing the sign of that bias is what lets the paper use the exact simulation as a reference for real systems, and it explains why the image network in Figure 2 needed twice the simulation’s dose of real data.

Figure S3. Re-baselining a collapsed population locks in the damage

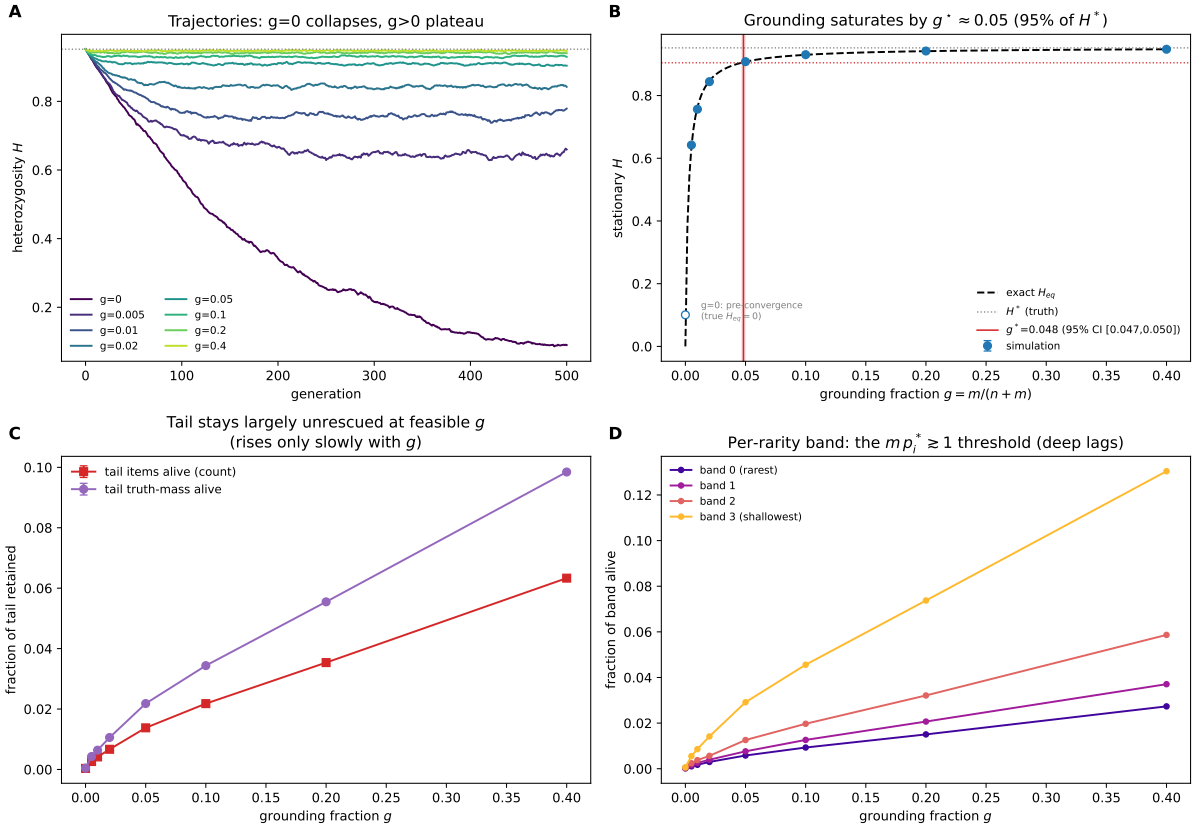


What was done. A tempting shortcut in practice is to declare a model’s current output the new “ground truth” and stop keeping the original data. The simulation tests what that does. Two hundred generations in, and again at 300, the population’s current frequencies are frozen as the new reference for grounding and the original truth is thrown away (it is kept only to measure against). Four arms: re-baseline while still healthy (green); re-baseline after collapse (red); the same, but only allowed when diversity is above 0.75 (blue); never re-baseline (grey).

What you see. (A) distance from the original truth against generation. The red arm jumps at each re-baselining and never comes back; the healthy arm shows small steps; the gated and the never arms coincide. (B) diversity, with the gate’s threshold as a dashed line.

What it means. This is Muller’s ratchet in a population of models: once the rare knowledge is gone from every copy, nothing downstream can rebuild it, and re-baselining after collapse makes the loss permanent. A simple rule (never re-baseline while diversity is low) prevents it. The lesson is that remedies must act while copies of the rare knowledge still exist somewhere.

Figure S4. The full grounding sweep in the simulation

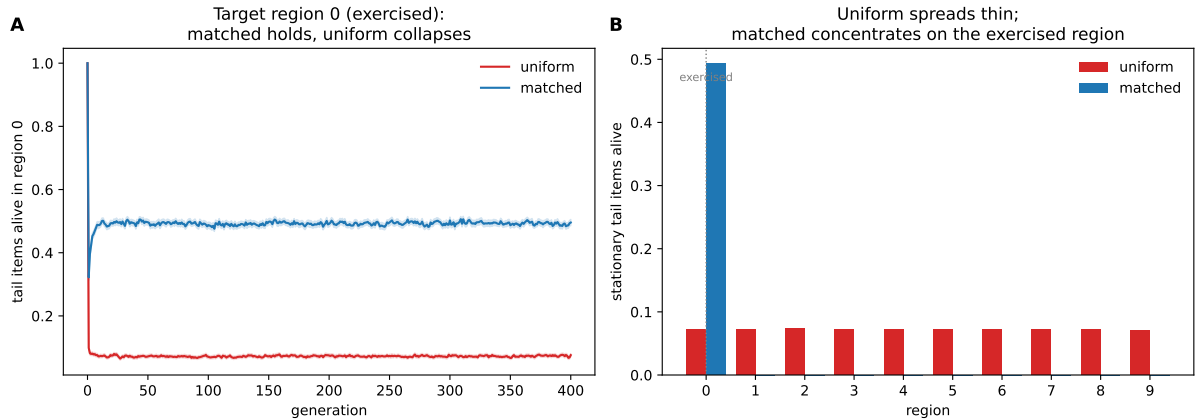


What was done. The complete version of the experiment summarised in Figure 2B: 1,000 knowledge items with a long tail of rare ones, 200 samples per generation, 500 generations, 100 lineages, and the fraction g of real data swept from 0 to 0.4.

What you see. (A) diversity over time, one line per g ; with no real data it declines steadily, with any real data it levels off. (B) the levelling-off value against g , with the exact prediction (dashed) and the real data's own diversity (dotted); the red line is the 95%-retention point at g of about 0.048. (C) the fraction of the rare tail that survives, counted by items and by their share of the truth; both rise with g but stay below 0.1 even at g of 0.4. (D) survival split into bands of rarity, from the rarest to the least rare; the rarest bands recover last.

What it means. Overall diversity is cheap to protect, but the rarest items are not. An item persists only when enough real examples of it arrive each generation, roughly one per generation, so protecting it costs about one over its frequency in real samples. The real-data budget is set by the rarest thing you refuse to lose.

Figure S5. Real data protects only the topics it covers

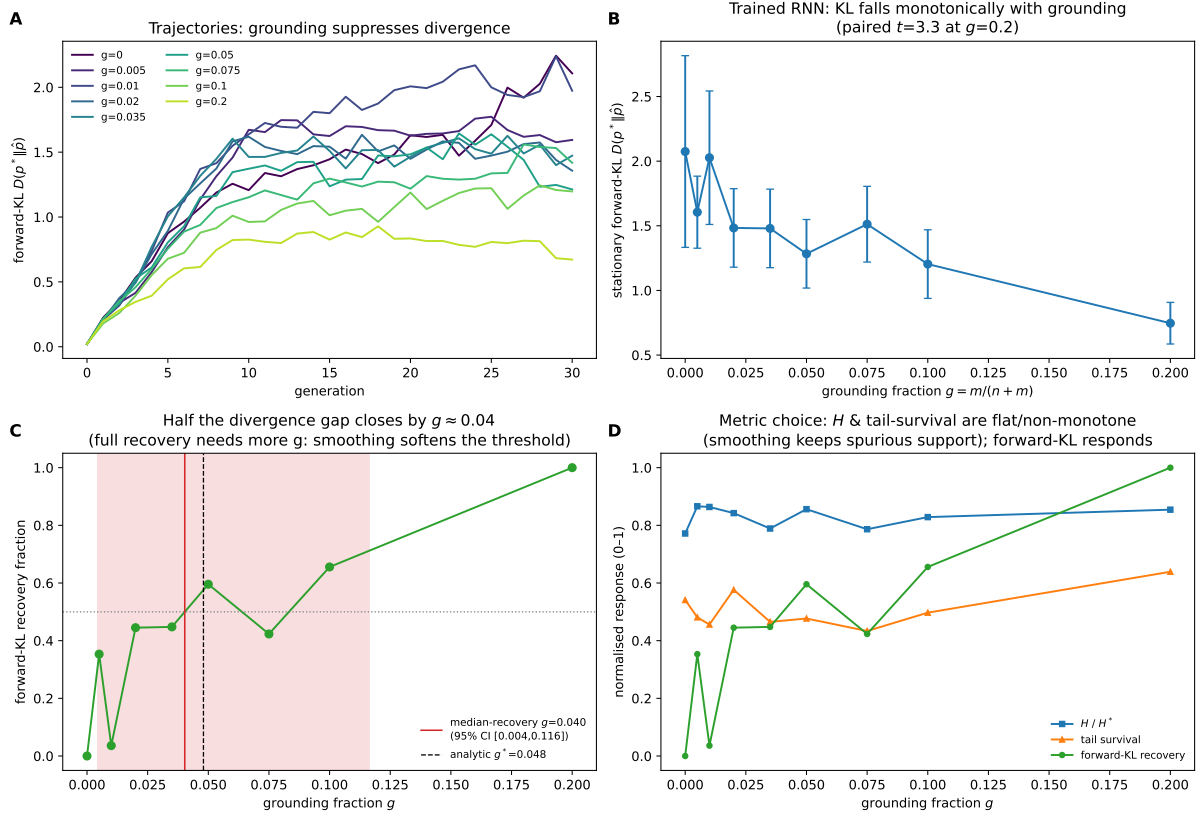


What was done. The 1,000 items were divided into ten topics. The same total budget of real data was spent in two ways: spread evenly over all ten topics, or concentrated on a single topic that the experimenter wants to protect.

What you see. (A) the fraction of that topic's rare items still alive, over 400 generations, when real data is aimed at it (blue) versus spread evenly (red). Aimed grounding holds about half the topic's rare items; spread grounding lets it fall to under a tenth. (B) survival per topic at the end. Aimed grounding protects its topic and leaves the others with nothing; spread grounding gives every topic the same low survival.

What it means. Grounding is not a general tonic. A fixed budget of real data protects the rare knowledge it actually contains, so it should be aimed at what matters, and targeting changes the cost of protecting a rare item substantially.

Figure S6. Grounding in a trained recurrent network

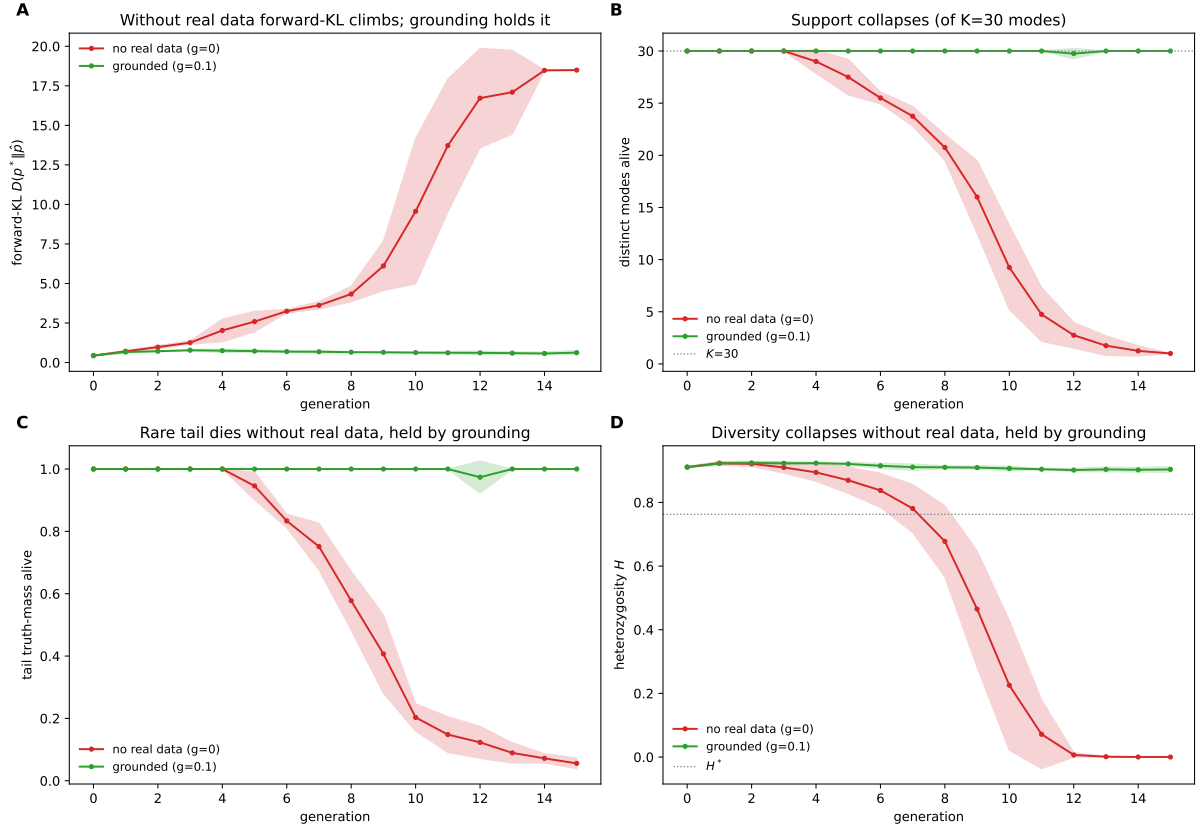


What was done. The grounding sweep of Figure S4 repeated in a trained recurrent network rather than the simulation: 256 kinds of item, 200 samples per generation, 30 generations, nine values of g from 0 to 0.2, eighteen repeats.

What you see. (A) distance from the truth over time; more real data suppresses the climb. (B) the final distance against g , falling steadily from about 2.1 with no real data to 0.75 at g of 0.2. (C) how much of the achievable improvement each g buys; half of it arrives by g of about 0.04, close to the simulation's 0.048, but the full improvement needs g near 0.19. (D) three ways of measuring collapse on one scale. Diversity is flat; the count of surviving rare items goes up and down with no pattern; distance from the truth improves cleanly.

What it means. The direction of the effect is the same as in the simulation, but the sharp threshold softens, and counting surviving items is the wrong ruler for a smoothing network, because it keeps inventing rare items that are not in the truth. Distance from the truth is the measure the paper uses for such networks.

Figure S7. Collapse and rescue on real handwritten digits, in numbers

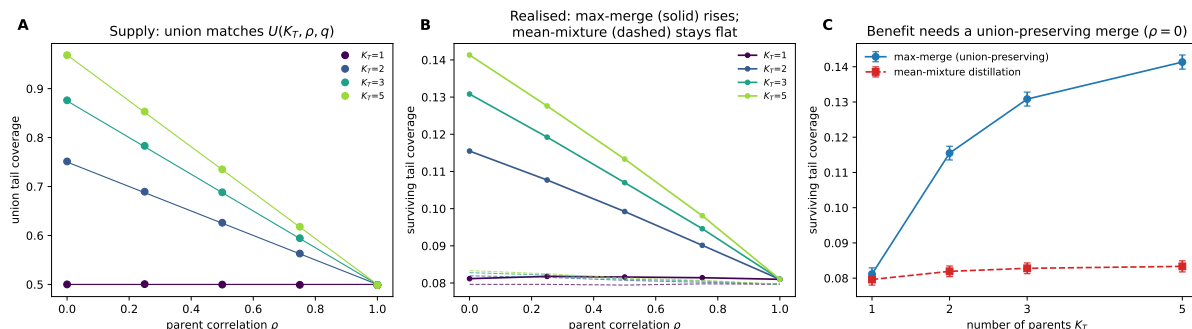


What was done. The experiment whose pictures are in Figure 2A, quantified. Thirty kinds of digit, a classifier reading the kind of each drawn digit with 98.5% accuracy, 6,000 drawings per generation, fifteen generations, four repeats, with 0% (red) or 10% (green) real digits mixed in.

What you see. (A) distance from the truth rises from about 0.5 to about 18 with no real data and stays near the floor with 10%. (B) the number of distinct kinds still being drawn falls from 30 to 1 without real data; with it, all 30 survive. (C) the share of the rare kinds still alive falls to 0.06 without real data. (D) diversity falls to zero without real data and stays near 0.9 with it.

What it means. Everything the simulation predicted appears on real images with an independent judge, and the dose of real data needed is about twice the simulation's, for the reason given in Figure S2.

Figure S8. Averaging parents cancels the benefit of having several; keeping the best of each does not

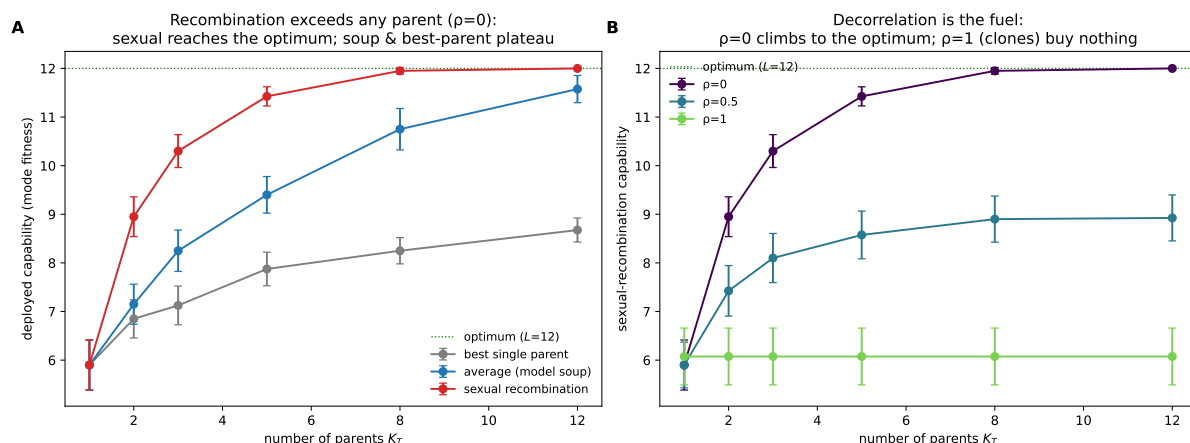


What was done. Several parent models each remember a random share of the rare items, and the experimenter controls how similar their shares are (from fully complementary to identical). A child is then built either by averaging the parents' output frequencies, or by keeping, for each item, the largest frequency any parent gives it (a *union*). The child then resamples, as every generation does, and the question is how many rare items survive in it.

What you see. (A) the fraction of the rare tail held by at least one parent, against parent similarity, one curve per number of parents; the lines are an exact formula and the points match it. (B) the fraction that survives in the child. Solid lines (union) rise with more and less similar parents; dashed lines (averaging) stay flat near 0.08 whatever the number of parents. (C) the same at zero similarity, against the number of parents.

What it means. Averaging dilutes each rare item by the number of parents, which exactly cancels the gain of having more parents to draw on: a conservation law. That is the AI form of *blending inheritance*, the pre-Mendelian idea that offspring are an average of their parents, which Fleeming Jenkin showed would swamp any rare favourable variant. Only an operator that keeps each parent's strongest contribution realises the benefit of several parents, and it needs a judge to say which parent that is.

Figure S9. Many complementary parents can produce an offspring better than any of them



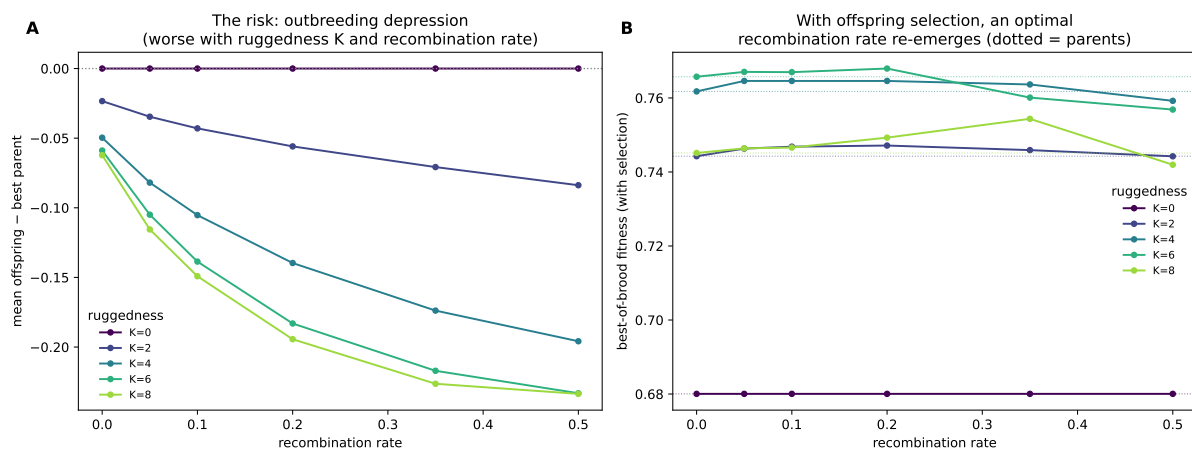
What was done. A capability is modelled as a string of twelve yes/no positions (a *genotype*

of twelve *loci*), and fitness is the number of positions that are right. Each parent is a specialist: confident and correct on the positions it has mastered, unsure elsewhere, and no parent has mastered them all. Offspring are built from 2 to 12 parents either by averaging or by taking, position by position, the answer of the parent most confident about it.

What you see. (A) fitness of the offspring against the number of parents, when parents master different positions. Position-wise recombination (red) reaches the perfect score of 12 with eight parents; the best single parent (grey) sits near 8.7; the average of parents (blue) reaches about 11.6 at twelve parents. (B) recombination against the number of parents when parents are complementary, half-overlapping, or identical clones; clones gain nothing.

What it means. This is the Fisher–Muller effect in its cleanest form: recombination assembles, in one offspring, good variants that arose in different individuals. Unlike biology, a model population is not limited to two parents, so the effect is unbounded. Figure 3B is this result in real language models.

Figure S10. When skills are entangled, blind recombination harms the offspring

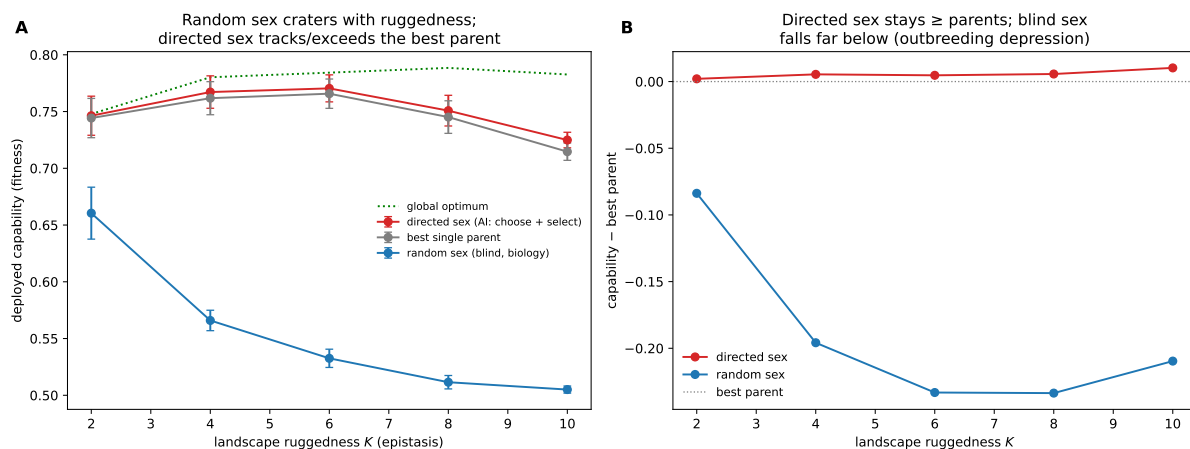


What was done. The same twelve-position genotypes, now on a *rugged* landscape (Kauffman’s NK model), in which the value of a position depends on what its neighbours hold, with a knob K from 0 (positions independent) to 8 (highly entangled). Parents are local optima found by hill-climbing, the model of a trained specialist. Offspring are made by recombining them at rates from 0 (copy a parent) to 0.5 (free shuffling).

What you see. (A) mean offspring fitness minus the best parent, against recombination rate, one curve per K . On a smooth landscape the difference is zero; as K grows the curves fall, more steeply at higher rates, down to about minus 0.23. (B) the fitness of the *best* offspring in a brood; on rugged landscapes it peaks at an intermediate recombination rate and falls back toward the parents under free shuffling.

What it means. This is *outbreeding depression*, well known in conservation biology: crossing two locally adapted populations can break up combinations of genes that only work together. The optimal amount of recombination shrinks as skills become more entangled. The design rule is to merge freely when skills are independent and sparingly, with selection, when they are not.

Figure S11. Directed sex: choosing and screening offspring rescues recombination

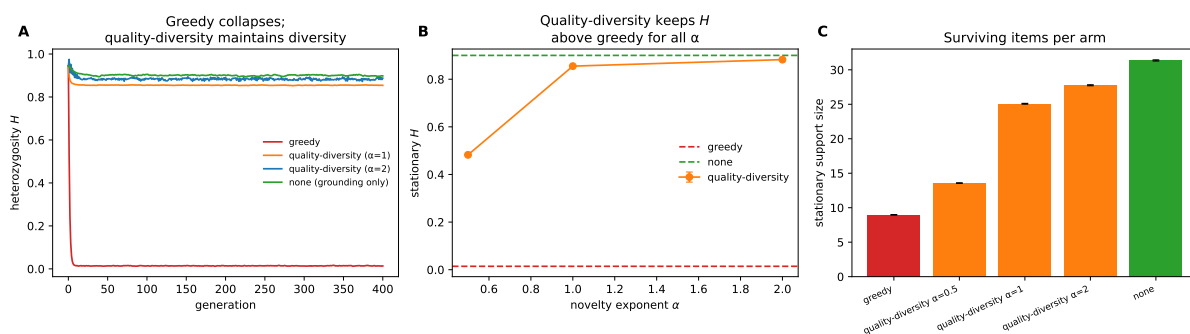


What was done. Biology is stuck with two random parents and no preview of the offspring. A model population is not: it can pick complementary parents, breed many candidate offspring, test them, keep the fittest and repeat. On the rugged landscapes of Figure S10 three strategies are compared: the best single parent (grey), random recombination (blue) and this *directed* recombination (red, five rounds).

What you see. (A) offspring fitness against ruggedness K , with the global optimum dotted. Random recombination falls from 0.66 to 0.51 as K rises; directed recombination tracks the best parent and the optimum at every K . (B) the same as a difference from the best parent; directed stays at or above zero, random falls to about minus 0.2.

What it means. The danger of Figure S10 is real but avoidable, by doing what no living population can. In language models this is “breed many merges, keep the best” (Figure 3 and Table S2), and it beat the single default merge in every seed on hard tasks.

Figure S12. Selecting for the best destroys diversity; rewarding novelty preserves it

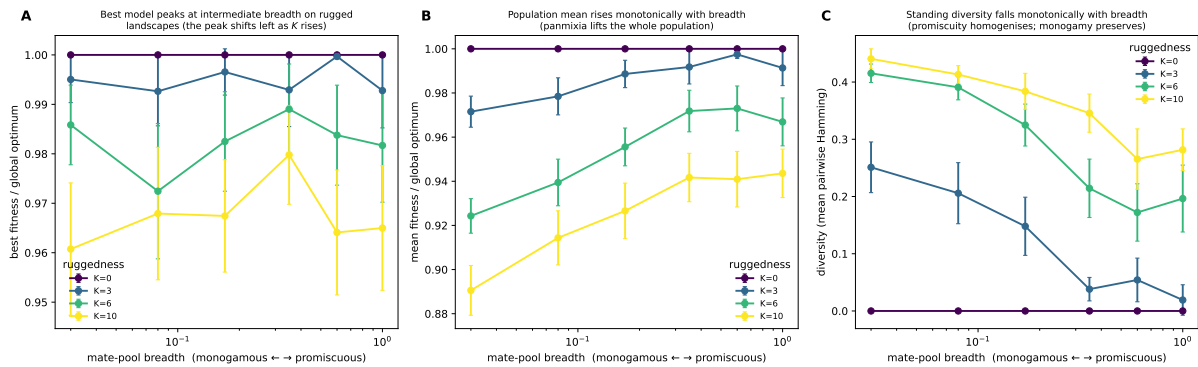


What was done. Each generation, the simulation now *selects* which items to keep, all arms receiving the same grounding. Three rules: no selection; *greedy*, keeping the items of highest true probability; and *quality-diversity*, which rewards an item for being rare as well as good, with a knob (alpha) for how much rarity counts.

What you see. (A) diversity over 400 generations. Greedy (red) collapses within a few generations to almost zero; quality-diversity at two settings and no selection hold a plateau above 0.85. (B) the settled diversity against alpha, rising from about 0.48 to about 0.88 as rarity is rewarded more. (C) the number of distinct items alive at the end: about 9 under greedy, 14 to 28 under quality-diversity, about 32 with no selection.

What it means. Chasing the best outputs is a fast route to collapse, because it is a directional pressure on top of drift. Diversity has to be an objective in its own right, since selection can only preserve variety that still exists. This is the “diversity preservation” ingredient of the composed society in Figure 4D to F.

Figure S13. Who should mate with whom: mating breadth on rugged landscapes

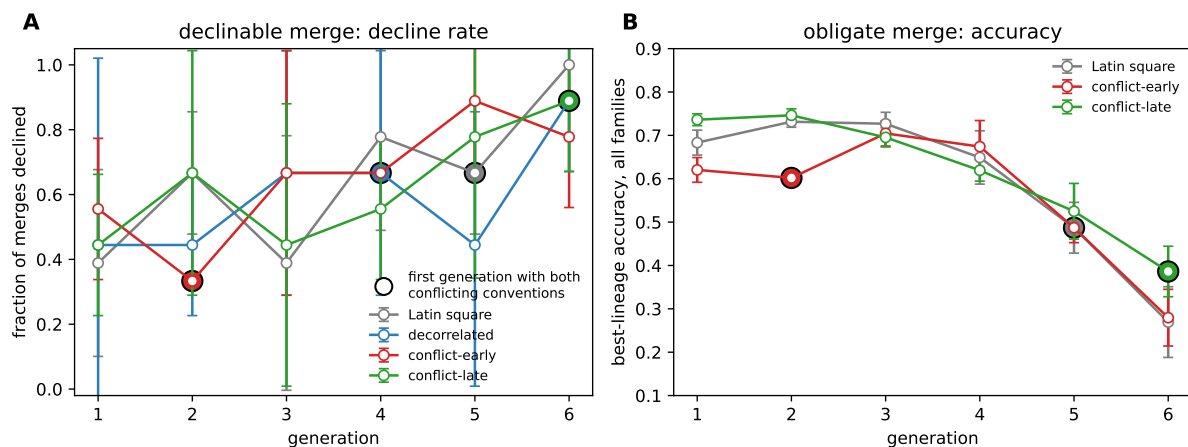


What was done. Forty-eight simulated agents sit on a ring. When an offspring is made, its second parent is drawn from a neighbourhood whose width is the knob: narrow (mating only with neighbours, like an isolated village) to the whole ring (anyone can mate with anyone). An offspring replaces the agent at its position only if it is fitter. Ruggedness K is swept from 0 to 10.

What you see. (A) the best fitness reached, relative to the optimum, against mating breadth, per K . On a smooth landscape every breadth reaches the optimum; as K rises the best breadth narrows (0.6 at K of 3, 0.35 at K of 6 and 10) and mating with everyone falls below it. (B) the population’s mean fitness rises with breadth at every K . (C) standing diversity (how different the agents are from one another) falls with breadth, fastest on rugged landscapes.

What it means. Wide mixing spreads a good variant fast but homogenises the population, so on entangled problems it loses the ability to explore several solutions in parallel. This is Sewall Wright’s classic argument for structured populations, reproduced here as a design rule: as skills become more entangled, keep merging local.

Figure S14. Do lineages stop merging because of conflict, or just because time passes?

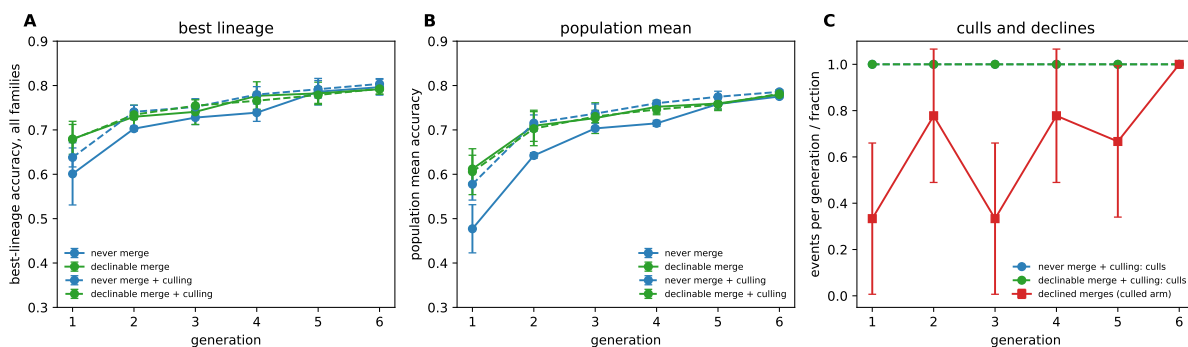


What was done. In Figure 4 three things rose together with generation: how long each adapter had been trained, how many skills it held, and the arrival of the two skills whose answer formats clash (one wants “yes/no”, the other “1/2”). To separate the third from the other two, two new syllabuses were run in which the clashing pair arrives either in the first two generations (conflict-early) or the last two (conflict-late), everything else rising exactly as before. Three seeds each, plus the two syllabuses from Figure 4.

What you see. (A) how often lineages refused a merge, by generation, for all four syllabuses; the filled marker on each curve is the first generation at which both clashing skills are present everywhere. Refusals rise with generation on the same schedule in all four. A statistical test that holds generation fixed finds no relationship between refusals and the presence of conflict, and a clear one with generation. (B) the accuracy of populations forced to merge every generation. The conflict-early population dips when the clash arrives, recovers, and then collapses from generation 5 like the others; the conflict-late population collapses from generation 4 before its clash has even arrived.

What it means. Moving the conflict by four generations did not move the collapse or the refusals. Conflicting conventions set how much damage each merge does, but something that grows with generation, adapter age or the number of skills carried, sets when the population can no longer repair the damage. Those two remain to be separated.

Figure S15. Survival of the fittest did not give merging lineages the edge

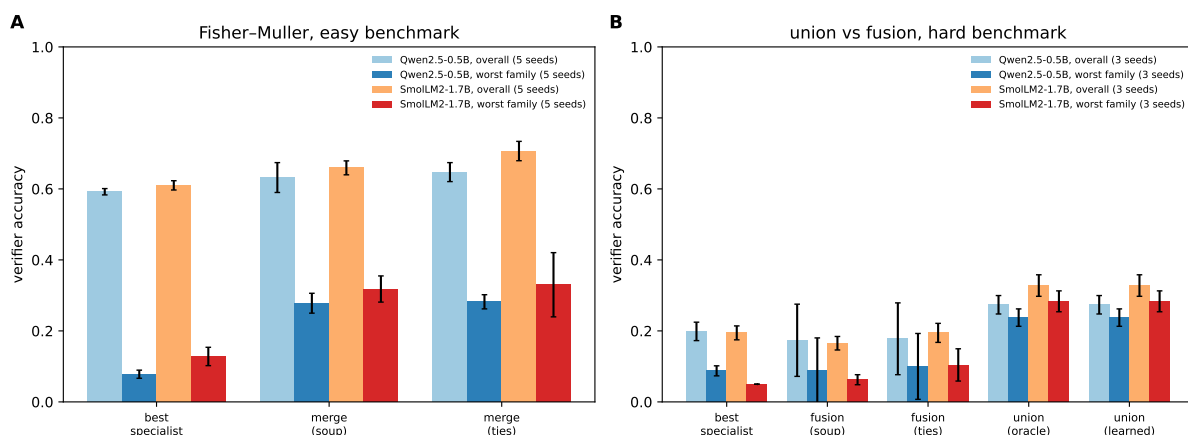


What was done. The population of Figure 4 never removed a lineage. Here, after each generation’s test, the worst-scoring lineage is deleted and replaced with a copy of the best (it keeps its own place in the syllabus). This is *differential reproduction*, the ingredient of natural selection the earlier population lacked. The prediction was that a lineage which assembles the skills first, by merging, would now leave more descendants and finish ahead. Never-merge and declinable-merge populations were run with and without this selection, three seeds each.

What you see. (A) best-lineage accuracy over the six generations for the four populations; the two selected ones are dashed. All four end within 0.01 of each other, near 0.80. (B) the average over the three lineages; selection lifts the average early (it copies the best genome into the worst slot), but the final averages converge too. (C) exactly one replacement happened every generation in every selected population, so selection was acting throughout.

What it means. Merging still bought speed, an early lead of about 0.08, and still bought no final advantage, with or without selection. Under a syllabus that eventually teaches every skill to every lineage, the ceiling is set by how much one adapter can hold, and both sex and selection only reach it sooner. The paper records this as a prediction it made and did not confirm.

Figure S16. The same results on a second, unrelated family of language models



What was done. Every language-model experiment in the paper used one family of base models (Qwen). To check that the two most-cited results are not peculiar to it, the experiments of Figure 3B and 3C were re-run, unchanged, on SmolLM2, a 1.7-billion-parameter model from a different laboratory with a different architecture and training data. Blue bars are the original Qwen runs, orange and red the SmolLM2 runs; lighter bars are overall accuracy, darker bars the worst task family.

What you see. (A) on the easy tasks the merged models beat the best single specialist on SmolLM2 in every one of five seeds, by about the same margins as on Qwen. (B) on the hard tasks, keeping specialists separate and routing questions to the right one beats averaging in every seed on both families, by a larger margin on SmolLM2, where the average even falls below the best single specialist.

What it means. The Fisher–Muller effect and the headroom rule hold on a second lineage of models. Results that depend on one model family are common in machine learning; these two do not.

Glossary

- **Adapter (LoRA).** A small set of extra trainable numbers added to a frozen base model, so that a specialist can be trained cheaply and two specialists can be merged by averaging their adapters.
- **Allele.** One of the alternative versions of a gene. In this paper, one of the alternative items of knowledge a model may hold.
- **Complementarity.** The share of a partner's skills that a lineage does not itself have. High early in the syllabus of Figure 4, zero at its end.
- **Epistasis.** When the effect of one gene depends on which other genes are present. For models: when a skill only works in combination with others (a rugged landscape).
- **Fisher–Muller effect.** The advantage of sex in bringing together, in one individual, beneficial variants that arose separately.
- **Fitness landscape.** A map from every possible genotype to its fitness. Smooth landscapes have one peak; rugged (NK) landscapes have many, so a population can get stuck on a poor one.
- **Forward KL divergence.** A measure of how badly a model covers the true distribution; it charges the model for every region where the truth has probability and the model has almost none.
- **Immigration–drift equilibrium.** The steady level of diversity a population settles at when new arrivals from outside balance the losses from drift. Its exact formula is the dashed line in Figure 2B.
- **Muller's ratchet.** In populations that never recombine, damage accumulates irreversibly, because once the best genome is lost it cannot be rebuilt.
- **Outbreeding depression.** Reduced fitness of offspring from parents that were each adapted to different conditions, because recombination breaks up combinations that only worked together.
- **Reproductive isolation.** The state in which two lineages can no longer produce viable hybrids; the defining boundary between species.
- **Routing.** Instead of merging specialists, keeping them separate and sending each question to the one that owns it.
- **Union (or max-merge).** Building a child by keeping, for each item, the strongest contribution any parent makes, rather than averaging the parents.
- **Wright–Fisher model.** The simplest mathematical model of a population: each generation is a random sample of fixed size drawn from the previous one. Its only force is chance, which is why rare variants disappear.