Pretraining History Leaves a Persistent Tail Prior in Small Market Forecasters

Forecasters that differ only in the era they are pretrained on, given identical later training and evaluated once on twenty sealed years, plus a set of retests of common claims about forgetting, plasticity and post training.

Balaji Daggupati Pebblebed November 2026 124k parameter transformers · 1,510 models · preregistered

ABSTRACT

We train small causal transformers from scratch to forecast the next day joint return distribution of the US market and twelve industry portfolios, using daily data from the Kenneth French data library. Models are identical except for the era they are pretrained on: the Depression and war years (from July 1926), the postwar boom (from 1946) or the stagflation years (from 1966), each cut to the same length. Every model then receives the same finetune on 1986 to 2001 and is evaluated once on a sealed test period from 2006 to August 2026 with frozen analysis code. On 1,062 calm test days, models pretrained on the Depression assign **1.30 times [1.18, 1.45]** the probability that boom pretrained models assign to a one day market drop of 3% or more (paired over 24 seeds, Holm adjusted p 4.8e−5), and **about half of the initial gap survives the finetune**. We report four results. First, the difference is a **tail feature rather than a level shift**: the ratio is near 1 at −1.5% and grows further out. Second, it fades slowly: the gap holds early in the finetune and about half remains at its end. Third, a second preregistered family confirms four of five lessons from development on the test years, including that late crisis data scars more than early data at a constant learning rate. Fourth, exploratory experiments indicate that plasticity methods leave the prior in place, while edits that move the weights far, including RL post training with GRPO, remove it. GARCH baselines beat every network on likelihood; the study measures training history, not forecasting skill.

A shorter, nontechnical essay, "Depression Babies," tells the same story for a general reader. This document is the fuller account: the design, the measurements, and the conditions under which each claim would be considered false.

01 Introduction

Pretraining data order, data mixing and continued pretraining all rest on an assumption that is rarely tested directly: that later training overwrites what a model learned first. In language the assumption is hard to examine, because a model's beliefs have no external answer key. Daily market data supplies one. A model that outputs a predictive distribution makes a statement that the next session grades; history provides pretraining regimes that differ sharply in their tail behavior; and the large modern crises fall after 2005, so they can be held out and read once. The human analogue is Malmendier and Nagel's finding that investors who lived through poor market returns take less equity risk for decades. The network version can be run as a controlled experiment with paired seeds.

This report asks whether that assumption holds for beliefs about rare events. The setting doubles as an instrument for several questions from model training, about data order, plasticity, unlearning and RL post

training, that lack an answer key elsewhere.

02 Design

2.1 Data

Inputs are daily value weighted returns for the market and the twelve industry portfolios in log percent, with real trading dates kept, including Saturday sessions before 1952.

2.2 Eras and controls

Each pretraining era is cut to the same number of trading days, 5,022. Every model then receives the same finetune, called schooling below.

TABLE 1 — PERIODS AND THEIR ROLES

Period	Dates	Role
Depression childhood	from 1926-07-01, fixed length	Pretraining
Boom childhood	1946 to 1965	Pretraining
Stagflation childhood	1966 to 1985	Pretraining
Shared schooling	1986 to 2001	Identical finetune for every model
Development	2002 to 2005	Tuning and exploratory work
Sealed test	2006-01-03 to 2026-08-31, 5,197 days	Read once, after the analysis was frozen

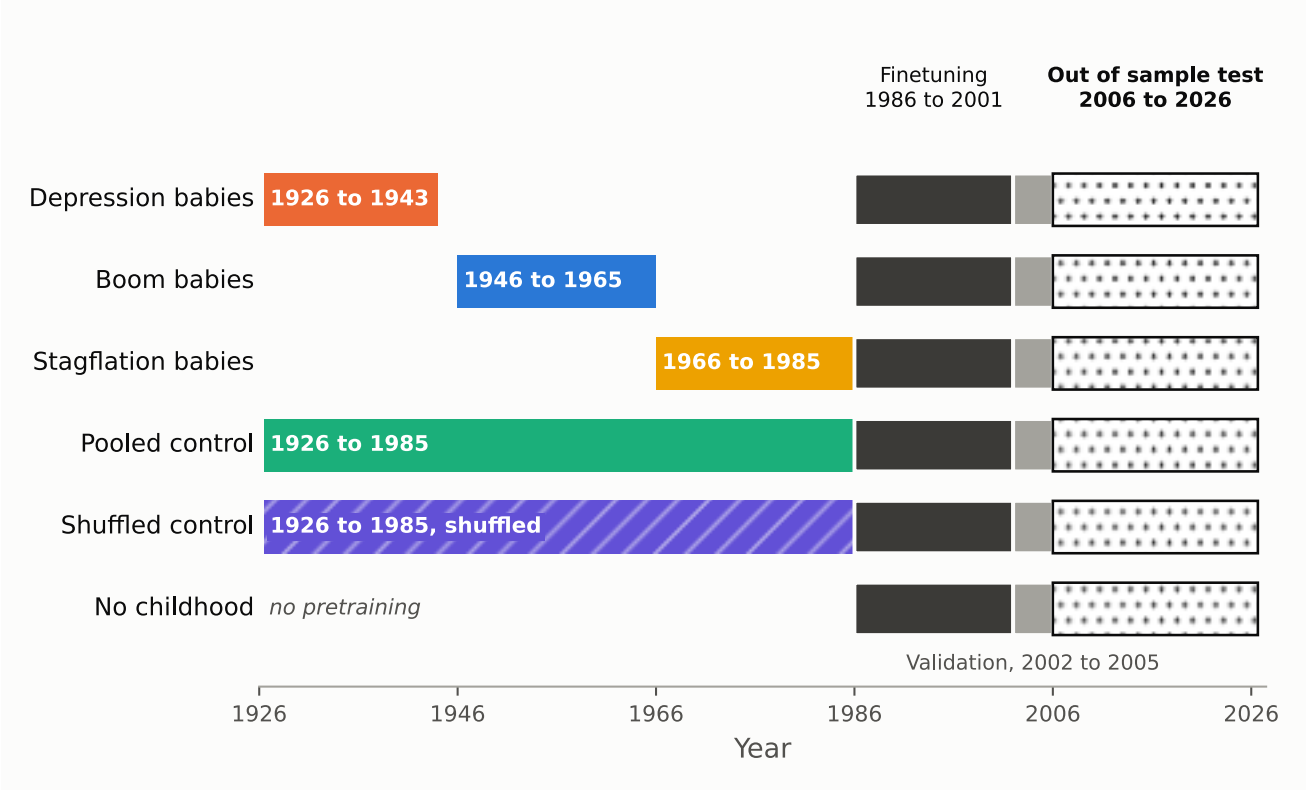


Figure 1. Study design. Every condition receives the same schooling; the test period is read once.

Three controls share the protocol: the 1926 to 1985 pool in shuffled blocks, the same pool in calendar order at matched compute, and no pretraining at all.

2.3 Calm days and the primary measure

We define a *scar* as a difference between two models' predictive beliefs that is attributable only to their pretraining data and that persists through identical later training. A day is calm if its trailing 20 day realized volatility is below the 25th percentile of that statistic over the finetune period, so the definition uses no future information. The primary quantity is *calm day fear*: the mean predicted probability of a daily market log return below a threshold τ, over calm days.

03 Model and Training

3.1 Representation

The model reads the previous 128 trading days of all thirteen series, plus the calendar gap since the previous session. Inputs and targets are divided by each series' trailing 60 day realized volatility, so an era can change the shape of a model's predicted distribution but not its overall scale.

3.2 Backbone and head

The trunk is a small causal transformer of about 124,000 parameters: width 48, 4 layers, 4 heads. The head outputs a mixture of four Gaussians over the thirteen next day returns, each a low rank (rank 2 plus diagonal) multivariate normal. The market marginal is a univariate mixture with the same weights, so tail probabilities are exact.

3.3 Objective and schedule

Training minimizes negative log likelihood of the next day returns. Schooling runs 2,500 steps with checkpoints every 250. Settings were chosen on a control's development likelihood and then frozen for every condition.

3.4 Paired seeds

Each Depression model has a boom twin that shares its initialization and the order of its examples, so the seed is the unit of replication; 24 paired seeds per condition enter the primary analysis.

04 Results

The primary family, analysis code and smallest effect of interest were frozen and timestamped before the test years were read. For each threshold the statistic is the paired difference in calm day fear between Depression and boom models at the final checkpoint, with a seed level bootstrap interval and an exact sign flip test, Holm corrected across six members.

4.1 Childhood survives schooling

TABLE 2 — PRIMARY PREREGISTERED FAMILY, TEST YEARS

Preregistered claim	Ratio [95%]	Holm p	Verdict
Depression models fear a –3% day more	1.30 [1.18, 1.45]	4.8e–5	supported
… a –1.5% day	1.04 [0.99, 1.08]	0.17	not supported (underpowered)
… a day below the schooling period's 1% quantile (–1.825%)	1.09 [1.03, 1.16]	0.012	supported
The gap survives late in schooling	48% of the initial gap remains	3.0e–5	supported
RL allocation policies built on them hold less equity	equity share 0.555 vs 0.553	1	not supported
More rollouts or a diverse ensemble close the gap	no	1	not supported

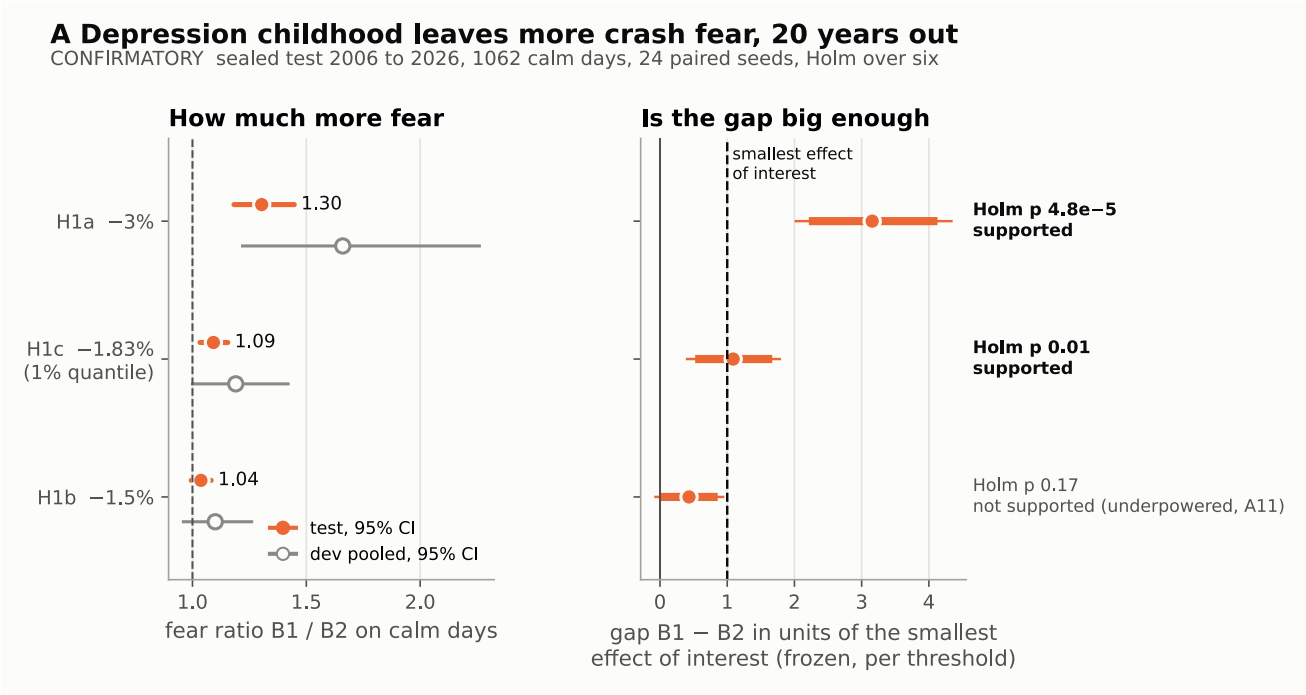


Figure 2. Calm day fear ratio at the three primary thresholds, sealed test beside the development estimate (development pooled ratio at –3% was 1.66 [1.22, 2.26]).

4.2 The scar is in the tail

Across seven preregistered sensitivity thresholds the ratio runs from 1.12 to 1.46, Depression above boom in every case and larger further out. The calm day excess kurtosis of the predictive distribution is 1.49 for Depression models against 0.96 for boom models. Both underpredict realized calm day tails: 5 of 1,062 calm test days fell below −3%, four to five times either model's expectation.

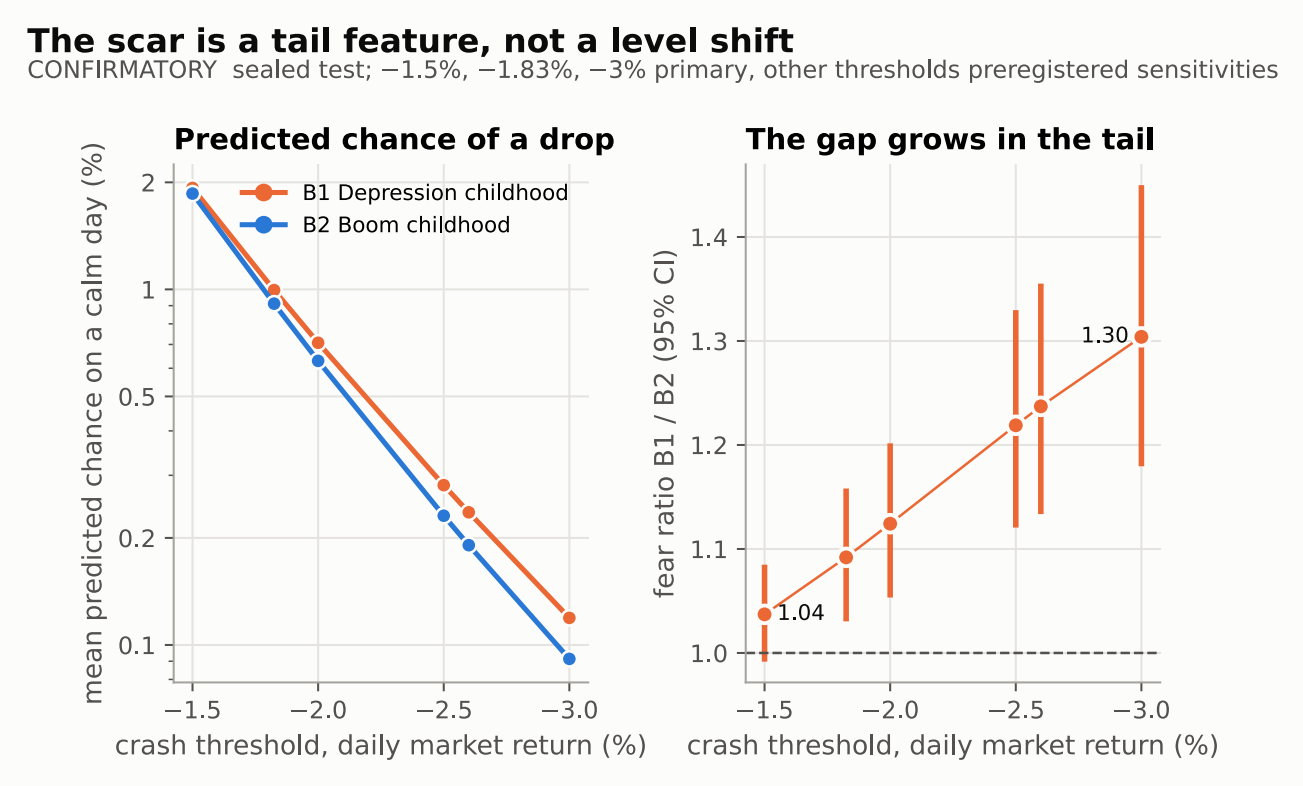


Figure 3. Calm day tail probability and the Depression over boom ratio across thresholds, sealed test.

4.3 It holds, drops, then fades slowly

The gap holds for about the first 500 steps of schooling, loses over a third by step 1,000, then keeps fading slowly; 48% remains when schooling ends. A descriptive exponential fit gives the decaying part a half life of roughly 970 steps above a floor too uncertain to estimate.

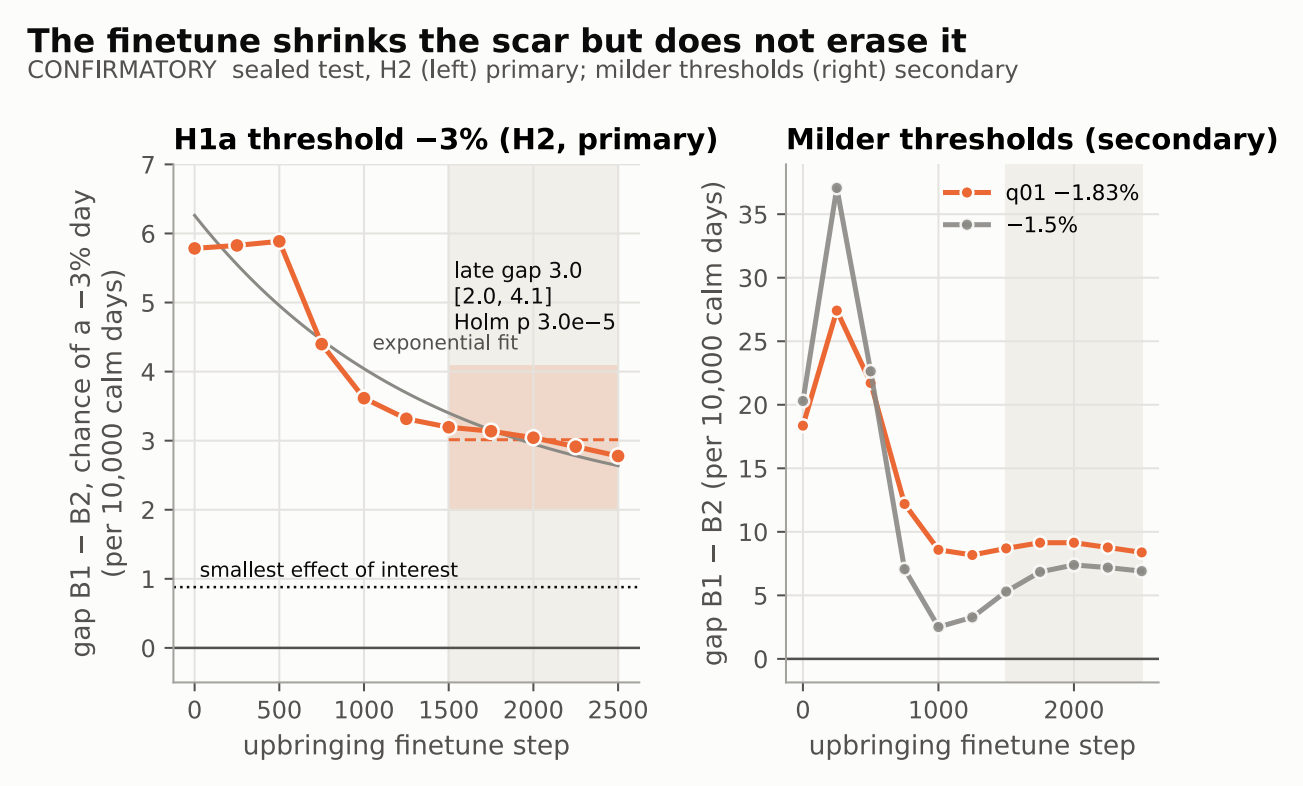


Figure 4. Depression minus boom calm day gap across schooling checkpoints, sealed test.

4.4 Controls

A model that saw the Depression returns only in shuffled blocks lands at 0.02 [−0.53, 0.47] on a scale where boom is 0 and Depression is 1, so it resembles the boom models; the calendar ordered pool lands at 0.40 [0.04, 0.67]. Stagflation models are the least fearful.

4.5 Behavior and baselines

Simple deterministic policies acting on the forecasts re-enter the market later after drawdowns when built on Depression models (2.3 against 1.0 days for a Kelly rule), while equity share does not differ in the predicted direction. On the test years GARCH t and GJR GARCH t beat every network on likelihood and CRPS, and even an exponentially weighted moving average beats every network on CRPS.

05 A Second Preregistered Family

Five lessons from development work were registered as a second family before unsealing and tested on the test years with models already trained.

TABLE 3 — SECOND PREREGISTERED FAMILY, TEST YEARS

Lesson	Estimate [95%]	Holm p	Verdict
Recency: late crisis exposure scars more than early, at constant learning rate	6.96e−4 [4.1e−4, 1.02e−3]	1.5e−4	holds
One bad day: a single spliced 1987-10-19 leaves a mark, at constant learning rate	8.2e−5 [4.8e−5, 1.14e−4]	1.5e−3	holds
Poisoning persists through schooling	1.42e−3 [1.21e−3, 1.65e−3]	1.5e−4	holds
Share of crisis data matters more than its placement	6.1e−4 [4.7e−4, 7.3e−4]	0.016	holds
Temperature scaling keeps the gap	−6.0e−5 [−1.45e−4, 1.9e−5]	0.17	fails

Four of five hold. The failure is temperature scaling, which grew the gap on the development years and shrank it on the test years.

06 Retests of Common Training Claims

Around the main study we retested widely held claims about training, each with a verdict rule committed before computation, on development data only. Of 43 scored claims, 7 survive a scorecard wide Holm correction and 23 a Benjamini Hochberg correction. A selection:

TABLE 4 — RETESTED CLAIMS, DEVELOPMENT YEARS

Claim	Verdict here
Plasticity methods (continual backprop, ReDo, plasticity injection, L2 toward init) let new data overwrite the prior	fails, all four methods (five settings)
Shrink and perturb removes the prior	holds, replicated
RL with verifiable rewards only sharpens the base model	fails: GRPO erases the prior on 20 of 20 seeds
A stronger KL penalty preserves the prior	partly: monotone, never complete
LoRA learns less and forgets less	fails, reversed at a higher learning rate
Quantization hurts tail knowledge first	fails, no change down to 3 bits
A calm cooldown erases the prior	fails, keeps 125% to 148%
Critical learning periods and primacy bias	fail; recency instead at constant learning rate; under cosine, data seen at low learning rate counts for less
Larger models forget more	withdrawn, did not replicate

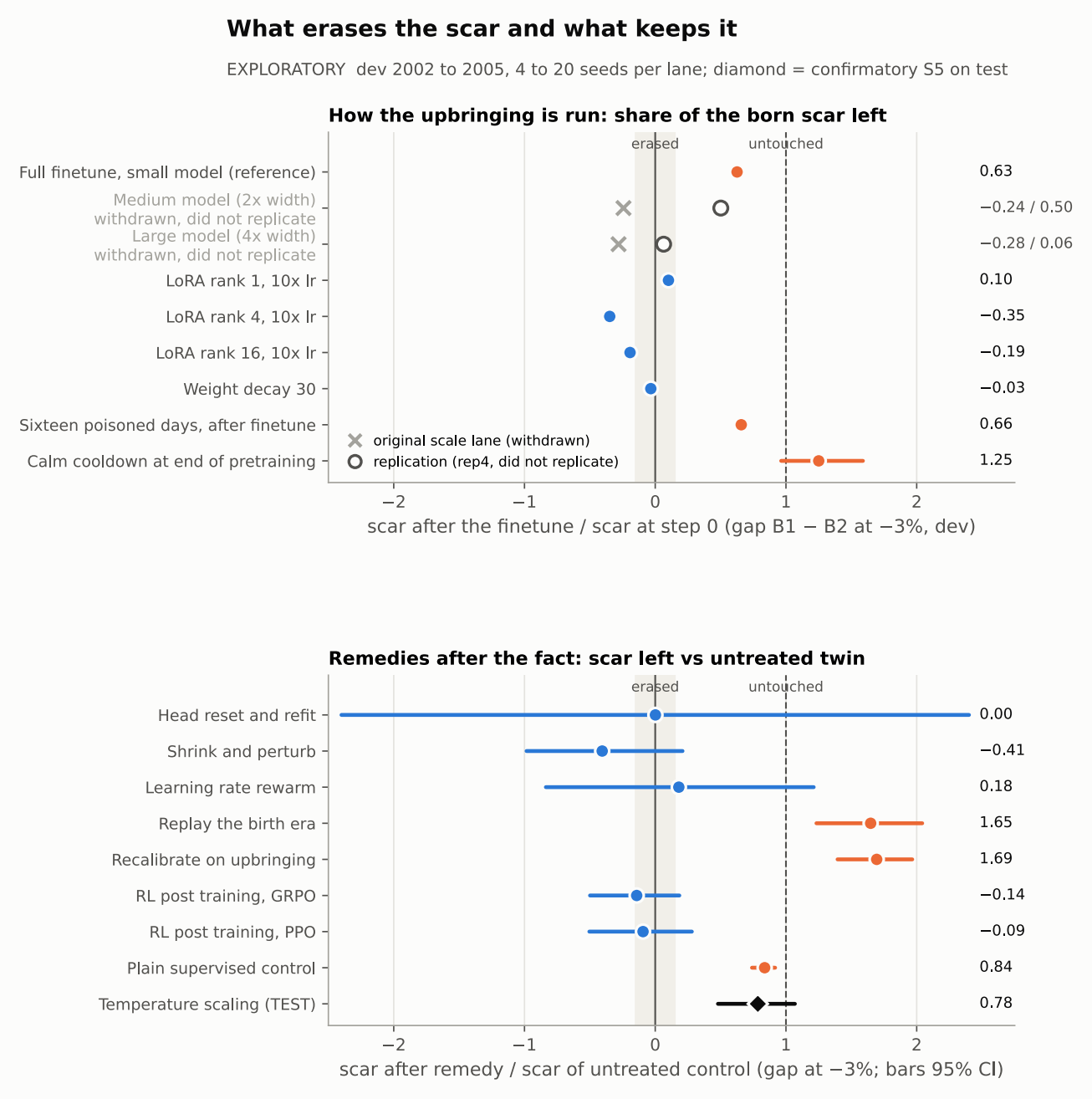


Figure 5. Share of the scar left after each intervention relative to its own control, development data.

Before unsealing, a synthetic answer key with known ground truth checked the instrument: a planted tail shape difference was detected and replicated on new data, a planted volatility level difference was normalized away by the scaled inputs as designed, and a null calibration with identical eras produced no false rejections.

07 Discussion

The result is small and durable, and the contributions are largely general; markets supply the answer key that language lacks. Persistence (Section 4.3) bears on any pipeline that assumes later stages wash out earlier ones: a standard finetune left about half of a pretraining belief in place. The order results (Sections

5 and 6) bear on data ordering and curricula: at a constant learning rate whatever came last left the bigger mark, while under cosine, data seen when the learning rate was already small barely registered. The remedy results (Section 6) bear on unlearning: methods that restore plasticity left the prior untouched, while edits that move the weights far, including GRPO, removed it at a cost to forecast quality. The through line is that, in this setting, pretraining history lives in the tails, where average metrics do not look.

08 Limitations and Falsification

All results come from one market, one small architecture family and one schooling recipe, and every interval reflects variation over training seeds on a single market history, not over alternative histories. The scar is small in absolute terms (about 0.3 additional expected crash days over 1,062 calm days), does not change RL allocation behavior, and the networks are weaker risk models than GARCH. The development estimate overstated the test effect, and two prominent development results did not replicate. We do not claim the effect transfers to large language models. The claim would be considered false if the gap vanished on another market with the same protocol, or if it disappeared at larger scale under identical schooling; both are open. Every verdict above was scored by a rule written down before the data it was scored on was read.

REFERENCES

1. U. Malmendier and S. Nagel. Depression Babies: Do Macroeconomic Experiences Affect Risk Taking? Quarterly Journal of Economics, 2011.
2. A. Achille, M. Rovere and S. Soatto. Critical Learning Periods in Deep Networks. ICLR, 2019.
3. S. Dohare et al. Loss of plasticity in deep continual learning. Nature, 2024.
4. J. Ash and R. Adams. On Warm Starting Neural Network Training. NeurIPS, 2020.
5. E. Nikishin et al. The Primacy Bias in Deep Reinforcement Learning. ICML, 2022.
6. D. Biderman et al. LoRA Learns Less and Forgets Less. 2024.
7. Z. Shao et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. 2024.
8. K. French. Data Library. Tuck School of Business, Dartmouth.