SCALE BUYS IMITATION, NOT THE DIVIDEND: POST-IMITATION RL ADDS A SIZE-INVARIANT GAIN IN AN IMPERFECT-INFORMATION GAME

Anonymous authors
Paper under double-blind review

ABSTRACT

Does the gain from reinforcement learning on top of an imitation model grow with the model? The answer decides whether compute goes into a larger base or into the RL stage, and it is rarely measured, because the gain is small against the variance of the tasks where it matters. We call the gain one RL stage adds over its own clone the *RL dividend* and measure it on a controlled ladder: five transformer clones (8M to 508M parameters) of one 1.4-billion-decision expert corpus in a four-player imperfect-information game (Riichi mahjong), one belief-state-distillation PPO recipe held fixed, a rules-accurate simulator, replicated runs, and reading rules fixed in advance. The dividend does not scale. From 22.5M to 155M parameters it is $+80$ to $+110$ Elo at every rung (25 PPO updates; 30 at 155M, which reads $+64$ at its 25th), replicated by an independent second run at two rungs, and the three post-RL policies finish within four points of even against each other on the same tables. One RL stage on the 22.5M model (about five GPU-hours on one consumer card, an hour of it PPO) beats the 155M clone on 59% $[56, 62]$ of shared tables; scaling the clone seven-fold and training it longer buys 55% $[52, 58]$, and scaling it from 22.5M to 56M buys nothing (49%). PPO does the same thing at every size (entropy collapses from 1.4 to 0.5 within five updates, 16–23% of held-out human decisions move, two independent 508M runs agree on 97.3% of decisions). At the ends of the range the dividend is run-dependent: 0, $+36$ and $+53$ Elo at 8M; $+45$, $+77$ and $+99$ at 508M, where two of the runs differ by 58 Elo on the same tables and one keeps rising to $+110$ with fifteen further updates. In two small imperfect-information games with exact metrics the dividend is bounded by the teacher's headroom, nil against a near-Nash teacher, growing with capacity against a near-deterministic exploitable one and growing less against a noisy one, the regime a human teacher is in; a ladder of language models pretrained on one corpus (Pythia, 14M to 410M), cloned from one noisy teacher on a synthetic task, takes the same dividend at every size. We also report how we got the shape wrong first: three silent infrastructure defects (stale trajectories from 384 games in flight, a value-head gate dropped in a refactor, mislabelled discard tokens in the deployed encoder) sat behind a spurious inverted-U with every training diagnostic inside its healthy band; two reproduced it on the repaired stack, one had shifted every Elo of the earlier one, and null readings on a fixed instrument and a code audit caught them. Single-run RL scaling curves are not evidence. Scale buys imitation; it does not buy the dividend.

1 INTRODUCTION

Reinforcement learning on top of an imitation model is now the stage that turns a clone into a deployed policy, in language models and in games alike, and the scaling question that matters for it is not the clone's: does the gain the RL stage adds grow with the model it is added to? If larger clones earn larger dividends, scale is the lever; if the dividend is constant, the RL stage is a fixed step that no amount of scale moves, and a small model with the step can meet a large one without it. The clone's own scaling is the easy part: its cross-entropy against expert play falls smoothly with parameters and, in single-agent games, its return follows the loss (Tuyls et al., 2024; Pearce et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2025). In a four-player imperfect-information game it follows grudgingly: our clones' strength rises to 8M parameters and then saturates near the expert's own entropy, so that a seven-fold larger clone wins 55% of its paired games against the smaller one (§3). Falling loss and a saturating gain in play is what scale buys imitation; we call the gain one RL stage adds over its own clone the *RL dividend*, and this paper measures it as a function of the clone's scale.

Our testbed is Riichi mahjong (three hidden hands, a stochastic wall), the regime where this question is hardest to answer honestly: a seat's final score has a standard deviation of about 12,400 points per game against about 55 points per Elo of true edge, so strength is expensive to measure while loss is free. The paper's first job is therefore an instrument: a rules-accurate simulator, a fixed evaluation field, paired same-table comparisons, two or three runs at four of the five rungs, and a reading rule fixed for every number before its data (Appendix A).

The argument. *First* (§4): under one PPO recipe at five sizes, the dividend over each rung's own clone is $+80$ to $+110$ Elo from 22.5M to 155M (25 PPO updates; 30 at 155M), replicated by a second run at two rungs, and the three post-RL policies finish within four points of even on the same tables (two pairs level, the third above by a hair). *Second* (§4): one RL stage on the 22.5M clone beats the 155M clone on 59% of shared tables after about five GPU-hours, where scaling the clone seven-fold and training it longer buys 55%. *Third* (§5): PPO's fingerprint on held-out human decisions is the same at every size, and two 508M runs agree on 97.3% of decisions. *Fourth* (§6): at the ends of the range, 8M and 508M, whether a run realises the dividend within the budget is run-dependent; two of the 508M runs differ by 58 Elo on the same tables while nothing in their training records or their decisions on human states separates them. *Fifth* (§7): in two small imperfect-information games with exact metrics the dividend is bounded by the teacher's headroom, and its capacity-dependence follows the teacher's noise: growing against a near-deterministic exploitable teacher, growing less against a noisy one, nil against a near-Nash one; a ladder of language models pretrained on one corpus (Pythia, 14M to 410M) cloned from one noisy teacher takes the same dividend at every size (Appendix F). *Sixth* (§8): on the stack we used first the same family gave an inverted-U; three silent defects sat behind it and behind the repaired stack's first pass, none visible in a training curve, and a dose-response experiment reproduces the first at 22.5M.

Contributions. (1) The RL dividend as a function of imitator scale on one recipe at five sizes, with pre-registered readings, replication runs and same-table comparisons: size-invariant from 22.5M to 155M, run-dependent at the ends. (2) One RL stage outweighs a seven-fold scale-up of imitation on the same tables. (3) A size-invariant behavioural fingerprint of PPO on held-out human decisions. (4) Two small games with exact metrics that locate the regime: the dividend is the teacher's headroom, and a noisy teacher weakens its dependence on capacity. (5) Three documented silent failures, two of which manufactured a scaling shape, with the discipline that caught them.

2 SETUP

Domain and corpus. Four-player Riichi mahjong: 47-way action space, about 130 decisions per seat per game; 1.3796B expert state–action pairs (2.6M Tenhou games), shared across rungs. Each behaviour-cloning (BC) rung streams single-pass to its own convergence plateau; scale is the designed axis and exposure co-varies with it (Appendix C).

The family. Seven BC rungs share one architecture recipe (pre-LN transformer, SwiGLU + RMSNorm + dual policy head; width and depth scaled): 1.04M, 3.40M, 8.02M, 22.50M, 56.55M, 155.01M and 508.06M parameters. The RL rungs are the five from 8M up. The 155M RL start is the family's 155M BC continued for 38k further steps (§8 explains why); the paper says so wherever the 155M rung appears.

The RL recipe, one for all sizes. Five stages: an offline privileged critic; belief-state distillation (a belief head reconstructing the hidden state from the observable history); distillation of the critic through the belief; PPO with the belief critic (BSD-RL), 25 updates of 128 games (30 at 155M, an inherited launcher setting; §4 quotes its 25th update) with an opponent mix of 34% the current policy, 33% the run's own earlier checkpoints and 33% a fixed external anchor (Mortal v4 (Equimchan, 2024)); and checkpoint selection by a pre-registered rule (below). Every RL flag is explicit and identical across rungs apart from the seed and the 155M update budget; the upstream stages differ in the critic's batch (768, 384, 384, 256 and 128 from 8M to 508M) and in one phase-2 switch, a

belief-confidence weighting of the distillation losses, on at 8M and 508M and off at the three middle rungs (§6 returns to it). The executed recipe of every run, read from its launch record, its logs and its checkpoints, is the *recipe truth table* of Appendix B (260 machine checks; the one failing check is the earlier stack's 22.5M run, whose launcher carries a 384-in-flight setting that entered it after the run's collection, §8). The heads' hidden widths are fixed across scales; their input width follows the trunk, so their parameter counts grow with it (11M of non-encoder parameters at 8M, 81M at 508M).

The fixed stack. All numbers come from a simulator audited against 20,000 human wins (99.99% scored exactly; the engine we used first mis-scored 22.9%, §8) and a training stack with four defects repaired (the encoder's river-event tokens, the collector's double recording, the opponent pool's never-enabled historical share, the stop checkpoint's index); the earlier stack's numbers appear only in §8, as the object of study.

Evaluation. Every rung is swept on one fixed instrument: a 12-way cohort seating the rung's own BC, its RL checkpoints at every fifth update of its budget (5 to 25; 5 to 30 at 155M, a 13-way cohort), the four other rungs' BCs as a fixed gauge, and two Mortal anchors; 3,000 games per sweep, seats randomised per game. Strength is Bradley–Terry MLE on finish order (Bradley & Terry, 1952), anchored at Mortal v4 $= 1542$, with game-clustered bootstrap intervals; the *dividend* of a rung is the Elo of its RL checkpoint minus the Elo of its own BC in the same cohort. **Peak selection is pre-registered**: the peak update is chosen on the first 1,500 games of a sweep and its dividend is reported on the last 1,500, so no number is selected on the games that report it. Direct comparisons are *paired place-ahead*: the share of tables seating both policies on which the first finishes ahead (ties one half), with a game bootstrap; "level" means the interval contains 50%. No pairwise statistic is pool-invariant in a non-transitive free-for-all ([Anon.-B]), so every direct claim is a paired same-table one.

Pre-registration. Every reading rule (peak selection, replication across seeds, continuation, the head-to-head verdicts, the small-game rules, the dose-response reading) was committed to the repository before the result it reads was produced, with one exception: the Leduc expected-value rule was committed 70 seconds after its confirmatory run started, with early results already in the log; the plan records every rule's commit time, and a second confirmatory run on fresh seeds was read under that rule (Appendix A, §7). Where a rule turned out to be the wrong instrument, the paper says so and reports the pre-registered reading beside the corrected one.

3 IMITATION SCALES; ITS SKILL SATURATES EARLY

How much strength does scale buy the clone? On the fixed instrument, little past 8M. One ten-seat cohort of 4,000 games seats the seven clones, both 155M clones and two Mortal anchors (Table 1). Strength rises from 1M to 8M (-97 to -22 Elo against the anchor; each neighbour pair "above" on the paired tables) and then does not rise with parameters: 22.5M is level with 8M (48.0% $[43.9,\ 52.1]$ of 535 shared tables), 56M level with 22.5M (49.4%), 508M level with 155M (46.8%, 50.0%). The two 155M clones stand 7–37 Elo above the rest and are the one lineage with about 10^8 decisions of exposure against 14–43M elsewhere (Appendix C); parameters and exposure are confounded at that rung, and we attribute the height to exposure with that caveat. The pre-registered sentence test passes as registered: the upper end of the clones' 22.5M-to-155M gap, $+25$ $[+8,\ +43]$ Elo, lies below the smallest plateau dividend of §4 (and below the 155M run's reading at its 25th update). Held-out top-1 agreement with human play says the same from the other side: the five clones agree with the human on 0.760, 0.759, 0.772, 0.776 and 0.781 of 32,195 held-out decisions from 8M to 508M, a rise of two points across $64\times$ the parameters (the distillation that produces each RL start lowers it by three to six points before PPO begins: 0.718, 0.704, 0.736, 0.724 and 0.730). We read this as imitation of a stochastic expert saturating near the expert's own entropy: a larger network fits the noise better, not the play.

4 THE RL DIVIDEND DOES NOT SCALE

The five-point curve. From 22.5M to 155M the dividend is $+80$ to $+110$ Elo at every rung (Table 2: every rung's pre-registered dividend on the fixed stack, with the further runs where they exist): $+85.7$ $[+54,\ +116]$ at 22.5M (second run $+89.6$ $[+59,\ +122]$), $+91.5$ $[+63,\ +123]$ at 56M (second

clone	parameters	Elo vs Mortal v4 [95%]	paired against the clone below
1M	1.04M	-97 $[-116, -80]$	—
3M	3.40M	-42 $[-58, -24]$	59.7% above
8M	8.02M	-22 $[-38, -5]$	54.7% above
22.5M	22.50M	-40 $[-57, -24]$	48.0% level
56M	56.55M	-48 $[-65, -31]$	49.4% level
155M (v9_pool)	155.01M	-11 $[-27, +5]$	59.5% above
155M (RL start)	155.01M	-15 $[-32, +2]$	53.5% level
508M	508.06M	-22 $[-38, -6]$	46.8% / 50.0% level

Table 1: The clones on one instrument: Bradley–Terry levels against Mortal v4 (game-clustered bootstrap) in a ten-seat cohort of 4,000 games, and paired place-ahead over the next smaller clone ("above" when the game-bootstrap interval clears 50%). Mortal small sits at -55.

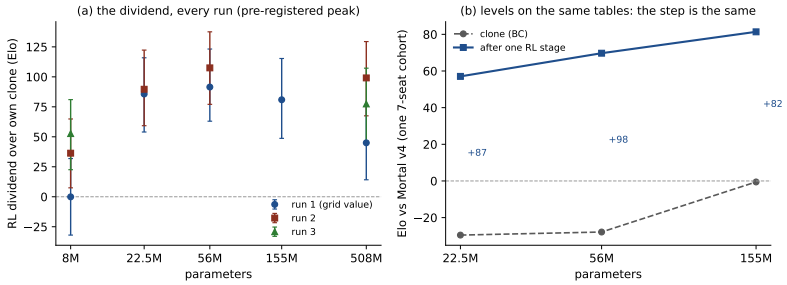


Figure 1: (a) The RL dividend over each rung's own clone on the fixed stack, every run, with game-clustered bootstrap intervals (peak chosen on the first half of a sweep, reported on the second). (b) Clone and post-RL peak against Mortal v4 in one seven-seat cohort of 3,000 games: the clone rises with size, the step on top of it does not.

run $+107.5$ $[+77, +138]$), $+80.9$ $[+49, +115]$ at 155M, at update 30 of a 30-update budget where the other rungs had 25: at its 25th update the same run reads $+63.7$ $[+28, +99]$, and on a 25-update grid the selection rule would have chosen update 20 and read $+44.4$ $[+12, +75]$. The two replication pairs differ by 4 and 16 Elo. At the ends the picture changes character rather than magnitude: 8M reads -0.1 $[-32, +32]$, $+36.3$ $[+7, +65]$ and $+52.7$ $[+23, +81]$ in three runs; 508M reads $+45.0$ $[+14, +76]$, $+99.1$ $[+67, +129]$ and $+77.4$ $[+46, +107]$ in three runs. §6 is about those two rungs; Figure 1 draws both halves.

The same tables put the three within four points of even. To compare rungs one must seat them together. A seven-seat cohort (the three peaks, their clones, Mortal v4; 3,000 games; about 870 shared tables per pair) gives 56M ahead of 22.5M on 51.7% $[48.6, 55.0]$, 56M ahead of 155M on 48.4% $[45.2, 51.8]$, and 155M ahead of 22.5M on 53.5% $[50.3, 56.6]$: two pairs read level by the rule of §2 and the third reads above by a hair; the three post-RL policies finish within four points of even against each other. Their Elo in that cohort rises with size ($+57$, $+70$, $+81$ against the anchor) mostly because the clones do (-30, -28, -1): the dividend on top is $+87$, $+98$, $+82$.

One RL stage against seven times the parameters. The same cohort seats the RL'd 22.5M model and the 155M clone. The small model with the RL stage beats the large clone on 59.0% $[55.7, 62.3]$ of 870 shared tables. The large clone beats the small clone on 55.1% $[51.8, 58.5]$. The RL'd 56M beats the 155M clone on 59.3% $[55.8, 62.5]$. Read as a ladder: from the 22.5M clone, one PPO stage is worth more on the same tables than multiplying the clone's parameters by seven and training it longer. The stage is cheap beside the clone: 54 and 61 minutes of PPO on one RTX 5090 in the two 22.5M runs, after about 3.5 hours of offline critic training (48,000 steps at 0.261 s), 18 minutes of belief training and 37 minutes of distillation, about 5.4 GPU-hours in all; the clone it beats streamed about 10^8 expert decisions (Appendix C).

More updates do not buy more. Both middle contenders end their grids still rising, so each was continued for 15 further updates with byte-identical flags and swept again. The continued checkpoints are paired against the grid peak on the tables seating both: 56M's u40 is ahead of its u25

rung	peak	run 1	run 2	run 3	zero-line verdict
8M	u25	-0.1 [-32, $+32$]	$+36.3$ [$+7$, $+65$]	$+52.7$ [$+23$, $+81$]	2 of 3 exclude zero
22.5M	u25	$+85.7$ [$+54$, $+116$]	$+89.6$ [$+59$, $+122$]	—	replicated
56M	u25	$+91.5$ [$+63$, $+123$]	$+107.5$ [$+77$, $+138$]	—	replicated
155M	u30	$+80.9$ [$+49$, $+115$]	—	—	—
508M	u20/u25/u20	$+45.0$ [$+14$, $+76$]	$+99.1$ [$+67$, $+129$]	$+77.4$ [$+46$, $+107$]	replicated

Table 2: The RL dividend in Elo over each rung's own clone on the fixed stack, with 95% game-clustered bootstrap intervals: peak chosen on the first 1,500 games of a 3,000-game sweep, reported on the last 1,500; the three 508M runs span 54 Elo. The 155M run's budget was 30 updates (its 25th reads $+63.7$ [$+28$, $+99$]); the first readings that the reruns at 22.5M and 155M superseded, -6.4 (384 games in flight) and $+13.8$ (a broken value-head gate), are in §8. Further runs re-seed only the trainer (deals and workers are independent draws); the zero-line verdict is the pre-registered replication rule (agree on whether the interval excludes zero; at 8M a third run decided by the two-of-three rule).

checkpoint	moved	KL(start‖ckpt)	entropy	human top-1	pon|offered	win|offered
8M, run 1 (0)	0.227	2.54	1.46 → 0.51	0.718 → 0.700	0.948 → 0.831	0.965 → 0.933
8M, run 2 (+36)	0.224	2.55	1.46 → 0.52	0.718 → 0.704	0.948 → 0.828	0.965 → 0.912
22.5M (+86)	0.192	1.91	1.44 → 0.54	0.704 → 0.740	0.943 → 0.777	0.951 → 0.911
56M (+92)	0.159	1.92	1.40 → 0.54	0.736 → 0.756	0.941 → 0.680	0.953 → 0.895
155M (+81, u30)	0.185	1.51	1.38 → 0.55	0.724 → 0.764	0.941 → 0.780	0.925 → 0.890
508M, run 1 (+45)	0.193	1.97	1.40 → 0.52	0.730 → 0.757	0.939 → 0.776	0.945 → 0.911
508M, run 2 (+99)	0.178	1.99	1.40 → 0.51	0.730 → 0.765	0.939 → 0.734	0.945 → 0.903

Table 3: PPO's fingerprint on 32,195 held-out human decisions at update 25 of each run (update 30 at 155M): the share of decisions whose argmax moved off the RL start, KL(start‖peak) in nats, entropy and human top-1 from the start to the peak, and the probability of pon and of taking a win when offered. Game-clustered bootstrap intervals are ± 0.011 or narrower on every entry except KL (± 0.05); the per-update curves are in the supplementary JSON.

on 54.3% [47.4, 61.8] of 173 shared tables and 155M's u45 ahead of its u30 on 53.3% [46.2, 60.9] of 169; both grid peaks stand by the rule; it does not say the dividend has reached a ceiling (the test resolves about ± 7 points, 56M's u40 reads $+33.5$ [$+5$, $+63$] above its u25 in the sweep, and 155M's peak is its 30th update).

5 What PPO does, at every size

Does a 508M policy change differently from an 8M one? On 32,195 held-out human decisions (60 games; the same frozen states for every checkpoint), it does not. Table 3 gives the fingerprint of seven of the eleven runs of Table 2 against their own RL start (the first run at every rung, and the second at 8M and 508M). By update 5 PPO has moved the argmax of 10–14% of the decisions and collapsed the policy's entropy from 1.38–1.46 nats to 0.49–0.56, where it stays; the moved share grows steadily to 16–23% by update 25–40 while the KL from the start reaches 1.5–2.6 nats. The probability of declaring riichi when offered stays at 0.96–0.99; the probability of calling pon falls from 0.94–0.95 to 0.68–0.83 (to 0.48 at 56M's update 40); the probability of taking a win when offered falls from 0.92–0.96 to 0.89–0.93. The one rung with its own signature is 8M: both of its probed runs, the one that gained nothing and the one that gained $+36$, move more decisions than any larger rung (0.22) and end *below* their start in human agreement (0.718 to 0.700 and 0.704) where every larger rung ends above it; the smallest student drifts furthest from human play, to no reliable gain. Beyond that, nothing in these curves separates the runs that gained from the ones that did not, nor the two probed 508M runs (seeds 42 and 43) from each other: their grid peaks choose the same action on 97.3% [97.1, 97.5] of the decisions and on 86.0% of the 6,190 decisions where either moved off the start, the same closeness as five and ten updates of one run's own drift (97.6%, 96.6%). Nor is the dividend bought with the belief: its wait readout sits at about a state-blind prior at every size, and PPO lowers it slightly (Appendix E).

6 THE ENDS: A RUN-DEPENDENT DIVIDEND

At 8M one run gained nothing and two gained $+36$ and $+53$; at 508M three runs gained $+45$, $+99$ and $+77$ (seeds 42, 43 and 44): the first stalled by update 20, the second kept climbing, the third rose smoothly to its plateau by update 20. Everything the training records can show says they are the same run twice. The first two 508M runs (seeds 42 and 43, the probed pair) share 57 explicit flags, the same start and the same number of games in flight; their per-update entropy, KL, clip fraction, losses and returns are indistinguishable; in weight space only the six modules trained from the distillation file move (54M parameters), together with an RL-only Monte-Carlo Q head (2.4M); the encoder, 473M, and the remaining heads, 25M, are frozen, 498M in all; the two runs' displacements of the six modules agree in norm to 3.1% pooled (7.9% per module) at every grid update and their directions agree with cosine 0.62 (policy head) to 0.93 (Q heads); and, as §5 showed, they agree on 97.3% of held-out human decisions. On the tables they are not the same run: in a six-seat cohort (both grid peaks, seed 42's u25, seed 43's u15, the shared clone, Mortal v4; 3,000 games; about 1,200 shared tables per pair) the second run's peak beats the first's on 58.0% $[55.2,\ 60.5]$, a Bradley–Terry difference of $+58$ $[+42,\ +72]$ Elo, the size of the gap between their sweeps. Continued for 15 further updates, the first run stays where it was (u30 paired against u20: 47.1% $[38.7,\ 53.9]$) and the second keeps rising (u40 against u25: 58.3% $[50.9,\ 65.7]$, to $+110.5$ $[+82,\ +141]$ over the clone). Nor is it in the critic: a pre-registered probe of the value baseline on the frozen held-out games (the mean squared change of V between an actor's consecutive decisions; log-loss against the realised placement) puts the stalled run's baseline at 1.29 times the noise of the other's at update 25 (0.130 against 0.101), short of the 1.5-fold margin the rule asked for, with log-loss 1.21 against 1.18; at 8M the two runs' baselines are alike (0.016 and 0.015). What the probe does show is a rung-level fact, of the three rungs probed after RL: both 508M runs' value baselines become 5.6 and 7.2 times noisier through PPO (0.018 at the start), where 56M's becomes 1.8 times noisier and 8M's 1.3 to 1.4 times. By the pre-registered two-of-three rule, 8M reads as a gain in most runs, smaller than the plateau's (-0.1, $+36.3$, $+52.7$), and stays run-dependent: the three values span 53 Elo where the plateau's pairs differ by 4 and 16. Half of that spread is the clone's own level moving between sweeps ($+3$, -21, -24 against the anchor); no same-table test separates the 8M runs.

What singles out the ends is not in the runs but in the family, and we can name three candidates. The 508M critic saw the fewest samples of the family before RL (5.1M against 18.4M at 22.5M and 56M), and the value baseline the probe reads (the clone's placement head, frozen in PPO and fed by the pool PPO trains) is the one PPO destabilises most of the three rungs probed. 8M is the smallest student, the slowest to realise its dividend in the small games (§7) and the one that drifts furthest from human play here (§5). And the two ends share one upstream setting the middle lacks: their distillation weighted its losses by the belief's confidence (a per-sample weight $\exp(-H_{\text{belief}})$), inherited from each rung's launcher and off at 22.5M, 56M and 155M; whether it matters is untested. We report these as candidates and do not choose between them; the paper's claim about the ends is the measured one: whether a run realises the dividend within the budget is run-dependent there, and it is not run-dependent in the middle.

7 TWO SMALL GAMES WITH EXACT METRICS

Is the flat dividend a property of RL or of this family? Two small imperfect-information games, run under the same opponent structure with exact exploitability and exact expected value by enumeration, bound the answer. In each game a *teacher* plays the role of the human corpus; a clone of it is trained at ten capacities ($h = 2$ to 1024 hidden units); PPO runs from each clone against the mix self 34% / own snapshots 33% / teacher 33% for 400 updates; the dividend is the expected value against the teacher per hand, maximised over the run's evaluation grid, minus the clone's; ten seeds per capacity. Three teachers per game: a near-Nash CFR strategy (mixed, no headroom), a transparent heuristic (exploitable, near-deterministic), and the same heuristic with 30% random actions (exploitable and noisy, the regime a human teacher is in). Figure 2 draws both games.

Leduc hold'em. Against the near-Nash teacher (exploitability 0.021) the dividend is nil at every capacity. Against the transparent heuristic (1.107) it is realised by at least nine of ten seeds at every capacity that fits the teacher (all ten at seven of the eight) and grows with capacity: $+0.21$ per hand at $h = 8$ to $+0.71$ at $h = 1024$. Against the noisy heuristic (0.988) it is realised everywhere and grows less: $+0.22$ at $h = 8$, then $+0.41$ to $+0.59$ from $h = 16$ up, $1.45\times$ from $h = 16$

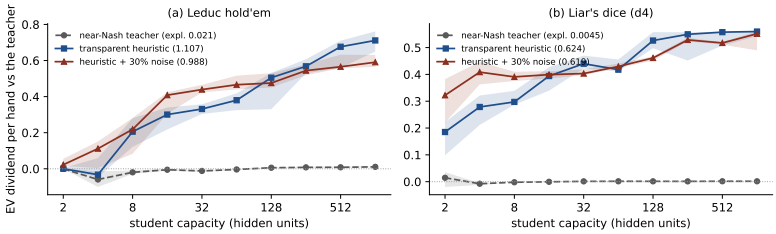


Figure 2: The two small games: the EV dividend against the teacher (median over ten seeds, interquartile band) as the student's capacity grows, for a near-Nash teacher (nil everywhere), a transparent heuristic (grows with capacity) and the same heuristic with 30% random actions (within the $\pm 25\%$ band in Liar's dice; $1.45\times$ over $64\times$ the capacity in Leduc, against $2.4\times$ for the transparent teacher).

to 1024 against $2.4\times$ for the transparent teacher; by the pre-registered shape rule it reads "size-dependent" (outside the $\pm 25\%$ band at $h = 8$, and its rise from $h = 16$, $+0.18$, falls short of the interquartile sum of 0.19 that the growth test asks for). Neither a ten-fold shorter budget nor an unexploitable field (the near-Nash strategy in the anchor's seat) removes the growth (both pre-registered predictions failed); with ten times the budget the small capacities come within the pre-registered tolerance of what the largest had at the short budget ($+0.42$ to $+0.61$ against $+0.71$) while the largest moves on ($+0.71 \rightarrow +0.85$): capacity buys speed, and whether it also buys the asymptote is unresolved (Appendix D). The pre-registered Leduc reading used exploitability as the dividend and returned nothing (a policy trained against a field containing its exploitable teacher moves toward a best response, away from Nash); Appendix D reports it beside the reading that answers our question. A second confirmatory run on fresh seeds (20–29) repeats every Leduc verdict but one: at $h = 8$, the smallest capacity that fits the transparent teacher, its dividend is realised by 5 of 10 seeds (Appendix D).

Liar's dice (one four-faced die each). The near-Nash teacher (0.0045) yields nothing. The transparent heuristic (0.624) yields a dividend that grows with capacity, $+0.28$ at $h = 4$ to $+0.56$ at $h = 1024$. The noisy heuristic (0.619) yields a dividend that is realised by every seed at every capacity that fits it and stays *within the pre-registered* $\pm 25\%$ *band*: $+0.39$, $+0.40$, $+0.40$, $+0.43$, $+0.46$, $+0.53$, $+0.52$, $+0.55$ per hand from $h = 8$ to 1024 (the band takes precedence in the rule; the rise from $h = 16$, $1.4\times$, would also pass the growth test, as the transparent teacher's $1.4\times$ from $h = 16$ does).

Three games, one rule: the dividend is the teacher's headroom, and its dependence on the student's capacity follows the teacher. A near-Nash teacher leaves nothing. A near-deterministic exploitable teacher leaves a best response that larger students represent better. A noisy exploitable teacher leaves headroom that small capacities take almost as fully as large ones (within the band in Liar's dice; $1.45\times$ against $2.4\times$ over $64\times$ the capacity in Leduc). The mahjong family's teacher is human play, noisy and improvable; the small games put that regime at the weak end of capacity-dependence, and mahjong's plateau is flatter still. A fourth testbed says the same in the language-model setting (Appendix F; its fit criterion was corrected twice, all three tabulated): on a ladder of models pretrained on one corpus (Pythia, 14M to 410M), cloned from one noisy teacher on a sorting task with exact reward, one RL recipe takes the same dividend at every size ($+0.31$ to $+0.34$ accuracy, five seeds each), a perfect teacher leaves only the clone's own sampling residue (0.02–0.04), and a teacher with one systematic flaw leaves a dividend that is run-dependent at four of the five sizes and unrealised at the fifth. The small games also show what they do not share with mahjong: at least nine of ten seeds realised the dividend at every capacity that fit the teacher (all ten in all but one cell), and the fresh-seed run repeats this except at Leduc's smallest fitting capacity (5 of 10); the run-dependence of §6 has a counterpart there only at that edge.

8 How we got the shape wrong first

The same family on the stack we used first, under the same recipe at four rungs and the production 155M lineage's own recipe at the fifth (its encoder unfrozen, a longer distillation, a different pool),

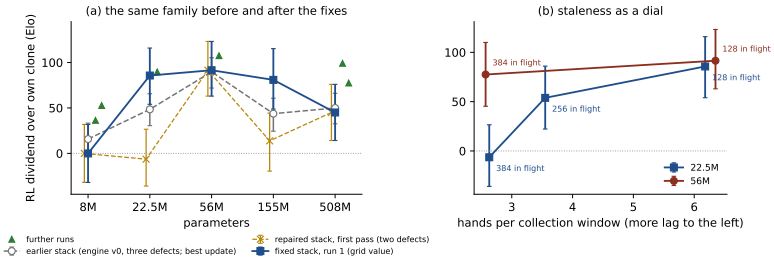


Figure 3: (a) The five-point dividend of the same family on the stack we used first (open circles: the spurious inverted-U; engine v0, three infrastructure defects, the 155M point from the production lineage's own recipe, each rung's best update on the games that report it), the repaired stack's first pass (crosses: two defects still present, the pre-registered peak rule) and the fixed stack (squares: the grid values; triangles: the further runs). (b) Staleness as a dial: the 22.5M dividend against hands per collection window at 384, 256 and 128 games in flight, and the 56M dividend at 384 and 128. Intervals are game-clustered bootstraps.

with each rung's best update chosen on the games that report it, gave a different answer: an inverted-U, $+16$ Elo at 8M, $+48$ at 22.5M, $+90$ at 56M, $+44$ at 155M, $+50$ at 508M, with every peak-versus-shoulder contrast significant. Three silent defects stand behind it and behind the repaired stack's first pre-registered pass (-0.1, -6.4, $+91.5$, $+13.8$, $+45.0$: an inverted-U by the plan's own rule; Figure 3a): two produced that first pass, the third had shifted every number of the earlier stack, and none was visible in a training curve.

Stale trajectories. The first fixed-stack 22.5M run read -6.4 $[-36, +27]$: no dividend. It differed from every other run in one setting outside the trainer's 57 printed flags, an environment variable of its launcher: the simulator held 384 games in flight instead of 128, so each trajectory spanned about three collection windows and reached PPO with log-probabilities of a policy three updates old (2.6 hands per collection window against 6.2–6.4 with 128 in flight). Rerun with 128 in flight and nothing else changed, the rung read $+85.7$ $[+54, +116]$ and its replication $+89.6$. A pre-registered dose-response at 22.5M makes staleness a dial rather than an accident (Figure 3b): with 256 games in flight (3.6 hands per collection window against 6.2 at 128 and 2.6 at 384) the same recipe reads $+53.9$ $[+22, +86]$ at update 25, having gained nothing through update 20, against $+85.7$ at 128 and -6.4 at 384; the two steps are -32 and -60 Elo at this budget by the pre-registered reading (at matched update 25 the 384 run reads $+28.3$ $[-1, +59]$: steps -32 and -26), and the gain that survives arrives later. At 56M the same three windows of lag (2.6 hands per collection window, as at 22.5M) leave the dividend in place: $+77.5$ $[+45, +110]$ at update 25 against $+91.5$ with 128 in flight, gaining from update 5 on. The pre-registered reading is that the effect is confined to 22.5M at this budget (one run per cell); we do not claim it generalises.

A dropped value-head gate. A refactor silently removed a gating term from the value head the day after the 155M BC was trained. Under the shared recipe the 155M rung read $+13.8$ $[-20, +44]$. A pre-registered probe on the frozen held-out games found its value baseline 96 times noisier than the noisiest of the other rungs' starts (§6 defines the probe); restoring the gate recovered part of it, and rebuilding the rung from a clone trained under the current code read $+80.9$ $[+49, +115]$.

Mislabelled river tokens. The deployed encoder emitted river-event tokens that did not match what the training data meant; every RL batch and every evaluation of the earlier stack carried it, and a code audit of the encoder against the data specification found it, not an evaluation. Together with a rules engine that mis-scored 22.9% of wins, it shifted every Elo the earlier stack produced; the repaired stack is the one all other sections use.

What caught them. Not the training diagnostics: entropy plateaus at 0.55–0.58 for every run, successful and failed alike, and returns, KL spikes and the anchor's share of seats do not separate them either. What caught the first two was a null reading on the fixed instrument that disagreed with a number we already believed, then a recipe truth table that read every run's executed configuration from its own records (the staleness) and a value-baseline probe with a reading of the code (the gate);

the third came from a code audit. Single-run RL scaling curves are not evidence: the two rungs we now call run-dependent would each have been called a shape on one run.

9 THREATS TO VALIDITY

One recipe. Every number is one PPO recipe's dividend at a fixed budget of 25 updates (30 at 155M; 40–45 where continued); another recipe could scale differently. *One data budget.* The clones saturate at this corpus's entropy; a larger corpus moves the clones and may move the dividend. *Two runs at 22.5M and 56M, one at 155M, three at the ends*: enough to call the middle replicated and the ends run-dependent, not enough to estimate the ends' distribution. *The 155M start* is the family's clone continued for 38k steps, disclosed wherever it appears (the BC ladder seats both 155M clones). *Cohort drift*: the same clone's absolute Elo moves by up to 33 Elo between cohorts, so every cross-rung statement is a paired one inside one cohort, and the five-point curve reports within-cohort dividends only. *The small games' exact metrics are exact for the games we chose*, which have neither mahjong's four seats nor its variance. *The clone ladder is level past 8M*, and a reader may take Table 1 to say that scale buys no imitation either, so that the family is teacher-limited throughout. Two things separate the clone's saturation from the dividend's: the clone's held-out agreement with human play keeps rising (§3) and its paired edge is real (55% at seven-fold with longer training, 59.5% for the lineage with more exposure), where the dividend's same-table pairs sit within four points of even; and the small games show the same split with exact metrics. The claim is the second; the paper makes no claim that clones scale well.

10 RELATED WORK

Loss laws (Kaplan et al., 2020; Hoffmann et al., 2022), scaling of imitation, multi-agent RL and board-game agents (Tuyls et al., 2024; Pearce et al., 2025; Neumann & Gros, 2022; Mandlekar et al., 2021; Jones, 2021; Ruoss et al., 2024), and the emergence claim and its critique (Wei et al., 2022; Schaeffer et al., 2023) motivate measuring the quantity a scaling claim is about rather than its proxy. Post-imitation RL in games (Silver et al., 2016; in this game, Suphx, Li et al., 2020, supervised learning on expert logs followed by RL) and RL post-training of language models (Ouyang et al., 2022; Stiennon et al., 2020) stack an RL stage on a clone; Stiennon et al. (2020) reported models trained from human feedback outperforming supervised models several times their size, a precedent for the cross-size statement of §4. For language models the scaling of the RL stage itself is now studied directly: across Qwen2.5 models from 0.5B to 72B, larger models learn faster from RL, with a saturating efficiency term (Tan et al., 2025); under one scalable recipe a larger model reaches a higher RL asymptote where most recipe choices move only the compute efficiency (Khatri et al., 2025); RL with verifiable rewards sharpens what the base already samples, so the base's headroom bounds the gain (Yue et al., 2025); larger policy models gain less from RLHF with a fixed reward model (Hou et al., 2024); and the closest ladder to ours, chess models from 5M to 1B pretrained on games, then fine-tuned and trained by RL on puzzles, finds post-RL performance predicted by the pretraining loss (Shen et al., 2026). Those studies measure performance after RL on a base whose own strength, or whose teacher, grows with size; the quantity here is the gain over a clone of one fixed teacher, held constant across sizes, and to our knowledge it has not been measured across a parameter range under one recipe with replication.

11 CONCLUSION

Across a $64\times$ range of parameters, one RL recipe adds the same step from 22.5M to 155M ($+80$ to $+110$ Elo over the clone after 25 updates, 30 at 155M; within four points of even on the same tables; replicated where repeated), a step that outweighs a seven-fold scale-up of the clone; at the ends it is run-dependent, and nothing short of the tables tells the runs apart. Two small games with exact metrics say the step is the teacher's headroom and that a noisy teacher weakens its dependence on capacity; the language-model reading, that larger bases learn faster from RL, meets ours on the headroom: an LLM's teacher is its own pretraining, which grows with the model; this family's teacher is one corpus at every size. Three silent defects showed us a different curve first. Scale buys imitation; it does not buy the dividend.

REPRODUCIBILITY STATEMENT

Every number in the paper is generated by a committed script from a committed artifact, and a committed checker re-derives every headline number and every summarising range from those artifacts against the source: the sweep game streams, the report JSONs read by the pre-registered readers (peak selection, replication, continuation, head-to-head, small games), the fingerprint and wait-readout probes, and the recipe truth table built from every run's launch record, logs and checkpoints. The supplementary material carries the dated pre-registration plan, the truth table, the readers, the simulator audit and the training-infrastructure fixes as diffs against the earlier stack; checkpoints and game streams are released at camera-ready.

AI USE STATEMENT

The authors set the research question, the model family and recipe, every experiment and its pre-registered reading rule, and the interpretation of every result. In this work, we used generative AI tools (a large language model assistant, Claude by Anthropic, under the authors' direction and review) for writing: drafting and editing the paper's text, the pre-registration plan and the supplementary notes; and as a coding assistant for the analysis, plotting, reader and experiment-queue scripts. Of the tasks with required disclosure this amounts to implementing methods (the scripts above, including those of the small-game and language-model pilots) and interpreting results (first drafts of the readings, each verified by the authors against the committed artifacts); the assistant also suggested candidate experiments and readings among which the authors chose (feedback on research methodology; refining hypotheses). We have not used generative AI tools to generate synthetic data sets (the synthetic tasks are produced by code), to develop theoretical models or conceptual frameworks, to assist with translation, to clean or reformat data sets, or for qualitative or thematic data analysis; formulating mathematical claims and writing or supporting proofs are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: creating and editing software code, creating scientific figures, drafting and editing the paper, and summarising literature. We have reviewed all AI-assisted work: every number is re-derived from its committed artifact by a committed checker, two independent review sessions checked the paper against the artifacts before submission, and the authors read every line. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

EleutherAI. Pythia model cards: pythia-14m, pythia-31m, pythia-70m, pythia-160m, pythia-410m. `https://huggingface.co/EleutherAI`, 2023. Revisions cf967c0a, e556ace2, a39f36b1, 50f5173d and 9879c9b5.

Equim-chan. Mortal: an open-source riichi mahjong AI. `https://github.com/Equim-chan/Mortal`, 2024. Model v4 weights (file date 2024-03-08), as distributed with the Akagi client, used as the fixed external anchor.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Zhenyu Hou, Pengfan Du, Yilin Niu, Zhengxiao Du, Aohan Zeng, Xiao Liu, Minlie Huang, Hongning Wang, Jie Tang, and Yuxiao Dong. Does RLHF scale? Exploring the impacts from data, model, and method. *arXiv preprint arXiv:2412.06000*, 2024.

Andy L Jones. Scaling scaling laws with board games. *arXiv preprint arXiv:2104.03113*, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for LLMs. *arXiv preprint arXiv:2510.13786*, 2025.

Junjie Li, Sotetsu Koyamada, Qiwei Ye, Guoqing Liu, Chao Wang, Ruihan Yang, Li Zhao, Tao Qin, Tie-Yan Liu, and Hsiao-Wuen Hon. Suphx: Mastering mahjong with deep reinforcement learning. *arXiv preprint arXiv:2003.13590*, 2020.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning*, 2021.

Oren Neumann and Claudius Gros. Scaling laws for a multi-agent reinforcement learning model. *arXiv preprint arXiv:2210.00849*, 2022.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Tim Pearce, Tabish Rashid, David Bignell, Raluca Georgescu, Sam Devlin, and Katja Hofmann. Scaling laws for pre-training agents and world models. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 48542–48562, 2025.

Anian Ruoss, Grégoire Delétang, Sourabh Medapati, Jordi Grau-Moya, Li Kevin Wenliang, Elliot Catt, John Reid, Cannada A. Lewis, Joel Veness, and Tim Genewein. Amortized planning with large-scale transformers: A case study on chess. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jingyan Shen, Ang Li, Salman Rahman, Yifan Sun, Micah Goldblum, Matus Telgarsky, and Pavel Izmailov. Understanding reasoning from pretraining to post-training. *arXiv preprint arXiv:2607.16097*, 2026.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Zelin Tan, Hejia Geng, Xiaohang Yu, Mulei Zhang, Guancheng Wan, Yifan Zhou, Qiang He, Xiangyuan Xue, Heng Zhou, Yutao Fan, Zhongzhi Li, Zaibin Zhang, Guibin Zhang, Chen Zhang, Zhenfei Yin, Philip Torr, and Lei Bai. Scaling behaviors of LLM reinforcement learning post-training: An empirical study in mathematical reasoning. *arXiv preprint arXiv:2509.25300*, 2025. Accepted at ACL 2026.

Jens Tuyls, Dhruv Madeka, Kari Torkkola, Dean P. Foster, Karthik R. Narasimhan, and Sham M. Kakade. Scaling laws for imitation learning in single-agent games. *Transactions on Machine Learning Research*, 2024.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research (TMLR)*, 2022.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025. NeurIPS 2025.

A PRE-REGISTRATION RECORD

The dated plan (`P3_PLAN_20260912.md`) with every reading rule and the result read under it is provided as supplementary material verbatim. Its headings carry hand-written times that run one to 51 minutes off the commits that introduced them, so the plan tabulates every rule's commit time, which is the binding one; six runs were launched within a minute before their rule's commit, and one rule, the Leduc expected-value reading, was committed 70 seconds after its confirmatory run started, when that run's log already held early results. The plan says so, and a second confirmatory run on fresh seeds is read under that rule in Appendix D. The record comes together with the recipe truth table and its builder, every reader named in this paper with the report it wrote, the per-game streams and round statistics of every sweep the paper reads, the frozen states of the critic probe, the sweep configurations, the code of the models, trainers, collector, evaluation harness and simulator, and the infrastructure and simulator fixes as diffs. Identifying strings and non-English working notes are rewritten by the packager, which re-reads every staged byte before writing the archive.

B RECIPE TRUTH TABLE

The executed recipe of every run, generated from launch records, logs, checkpoints and git (260 checks; the one failing check is the earlier stack's 22.5M run, whose launcher file carries a 384-in-flight setting that entered it after the run's collection, the setting the first fixed-stack 22.5M run inherited, §8), is provided as supplementary material.

C LADDER GENEALOGY

Table 4: The family: exact parameter counts (one rule: the full BC network including auxiliary heads, without the pool layer, which adds 0.6–1.4% to the RL bases), BC lineage, BC steps, and the expert decisions each clone streamed (steps $\times$ batch 512, from the training logs; the 155M row counts its two segments under this family's code, on a production lineage that had streamed about 10^8). Three disclosed 508M irregularities in its BC: an LR stall and restart, a data-subset change at 54k steps, and an endpoint that is the uniform average of 16 checkpoints over a stable window of its pool continuation rather than one checkpoint. The 155M RL start is the family's 155M BC continued for 38k steps under the current code (§8); the earlier 155M BC is seated in the BC ladder of §3 beside it.

rung	preset	parameters	BC lineage	BC steps	decisions streamed
1M	p3_tiny	1,044,155	human → +pool continuation	18k+10k	14.3M
3M	p3_small	3,397,883	human → +pool continuation	18k+10k	14.3M
8M	p3_haku	8,021,307	human 30k → +pool 10k	30k+10k	20.5M
22.5M	p3_large	22,502,843	human 30k → +pool 10k	30k+10k	20.5M
56M	p3_chun	56,549,947	all-pool (22k → 36k)	36k	18.4M
155M	hatsu	155,013,947	v9_pool production BC, then 38k steps under the current code	([Anon.-A])+32k+38k	16.4M+19.5M on $\approx 10^8$
508M	p3_daisangen	508,064,059	all-pool; disclosed LR and data history	46k+38k	43.0M

D SMALL GAMES: PROTOCOL, FIGURES AND THE PRE-REGISTERED EXPLOITABILITY READING

Protocol (both games). Clones: a two-layer MLP encoder of width h with policy, value and Q heads, trained by reach-weighted soft cross-entropy on the enumerated teacher-self-play distribution (4,000 full-batch steps, Adam, lr 10^{-2}). PPO from the clone (value and Q heads fresh): clip 0.2, 4 epochs, minibatch 2,048, entropy 0.01, value 0.5, Q 0.5, lr 3×10^{-4}, $\gamma = 0.99$, $\lambda = 0.95$, 400 updates of 128 episodes. Opponent mix per episode: the current policy 34%, a snapshot of the run

(every 40 updates, pool of the last four) 33%, the fixed external opponent 33% (the teacher unless stated); only the learner's seat is trained on. Exact exploitability and exact expected value against the teacher by full-tree enumeration every 20 updates (every 40 in Liar's dice). Ten seeds per capacity, $h \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$. Leduc: the six-card two-round game; teacher A is the CFR average strategy (3,000 iterations), B a heuristic that raises with a king before the board and with a pair after it, folds a losing card into a raise, and calls otherwise (98% on its rule), C the same with 30% uniform actions. Liar's dice: one four-faced die each, bids (q, f) with $q \in \{1, 2\}$ increasing in (q, f) order, a *liar* call once a bid exists, ± 1 payoffs; teacher A the CFR average strategy (1,000 iterations), B a heuristic that bids its own face and calls at quantity two (98%), C the same with 30% uniform actions.

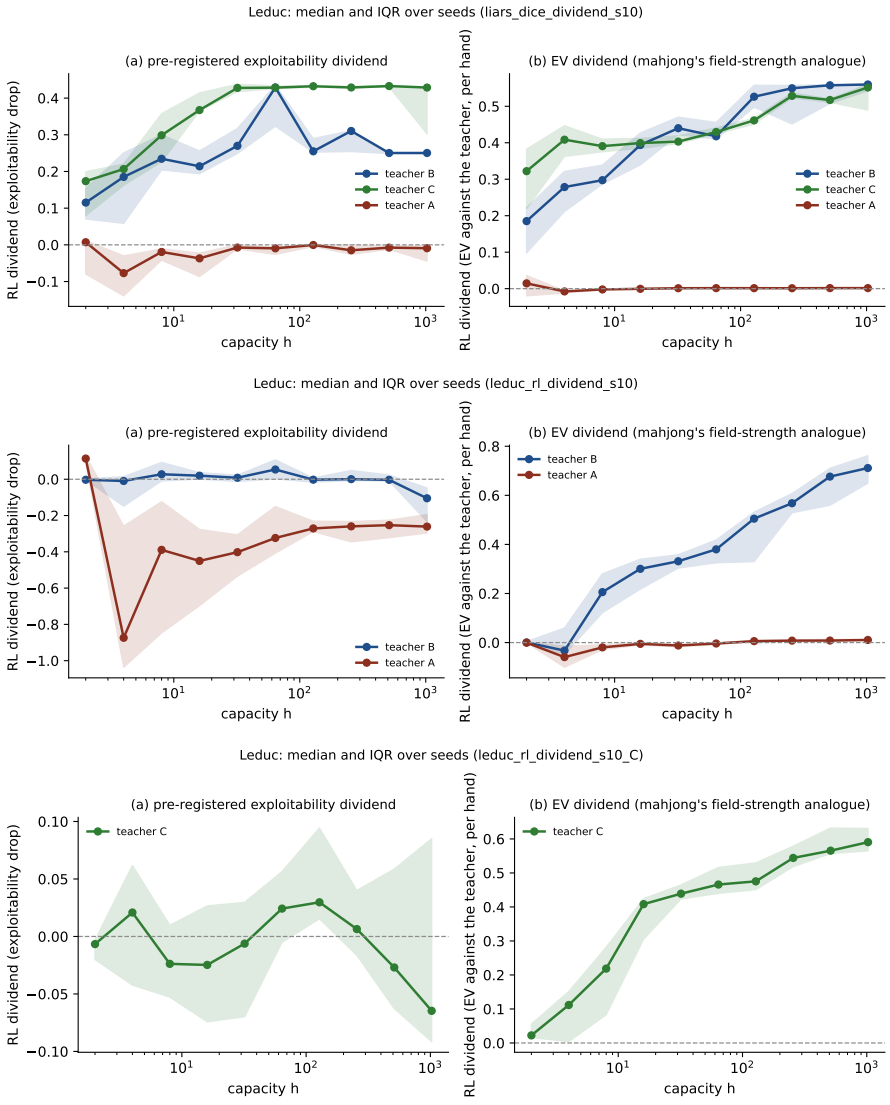


Figure 4: Small games, median and interquartile range over ten seeds per capacity. Top: Liar's dice, teachers B, C and A. Middle: Leduc, teachers B and A. Bottom: Leduc, teacher C. Left panels: the pre-registered exploitability dividend; right panels: the expected-value dividend against the teacher, the quantity the mahjong dividend corresponds to.

The pre-registered exploitability reading. The Leduc run was pre-registered with the exploitability drop as its dividend. It returned nothing: for teacher B the drop is about zero at every capacity that fits the teacher, and its final-checkpoint values are large and negative; for teacher A it is neg-

ative from $h = 8$. The run's own curves show why: from $h = 64$ up, the expected value against the teacher reaches $+0.33$ to $+0.45$ per hand within 100 updates (medians over seeds of the best value by then) while exploitability climbs from 1.1 to 5–8 and then relaxes as the historical pool fills; at $h \leq 32$ both move less. A policy trained against a field that contains its exploitable teacher moves toward a best response and away from Nash; the metric answered a different question. The expected-value reading was then pre-registered and confirmed on fresh seeds (10–19), which are the seeds every number in §7 comes from. Because that rule was committed 70 seconds after the 10–19 run started (Appendix A), a second confirmatory run on seeds 20–29 was read under it after an independent check: teacher B reads "grows with capacity" again (medians $+0.01$ at $h = 8$, then $+0.28$ to $+0.72$ from $h = 16$ to 1024; rise from $h = 16$ $+0.44$ against an interquartile sum of 0.26) but is realised at every plateau capacity only from $h = 16$ (5 of 10 seeds positive at $h = 8$, where seeds 10–19 had 10 of 10); teacher A is nil (medians within ± 0.02) and the headroom sentence is licensed again; teacher C reads "size-dependent" again (medians $+0.15$ at $h = 8$, $+0.34$ to $+0.61$ from $h = 16$ up), realised at every plateau capacity. In Liar's dice the exploitability dividend is positive at every capacity for teachers B and C, so there the two readings agree in sign.

Factors and budget (Leduc, teacher B, seeds 10–19). Neither of the two pre-registered predictions held: with the near-Nash strategy in the fixed opponent's seat the grand median dividend against B falls by 0.08 (against an interquartile sum of 0.23) and its medians still rise, though by the shape rule that cell reads "size-dependent" (two inversions); at 40 updates it still rises with capacity, and the 400-versus-40 gap exceeds the interquartile sum at six of eight capacities. With 4,000 updates (evaluated every 100; capacities 8, 16, 32, 64 and 1024) the small capacities come within the interquartile tolerance of the largest capacity's 400-update value ($+0.42$, $+0.50$, $+0.50$ and $+0.61$ per hand at $h = 8$, 16, 32 and 64 against $+0.71$) but the largest has moved on to $+0.85$, and $h = 32$ and 64 are still rising; the pre-registered sentence that fired is "capacity buys speed, not the asymptote", and whether capacity also buys the asymptote is unresolved at these budgets.

E THE BELIEF'S WAIT READOUT

The recipe's belief head reads the opponents' waits at a state-blind per-tile prior on the earlier stack, and PPO lowered it further. On the fixed stack the passage keeps its direction and loses most of its size: pre-RL to deployed wait AUPRC (4,096 states, base rate 0.048, prior 0.063) goes $0.0667 \rightarrow 0.0629$ at 8M, $0.0670 \rightarrow 0.0658$ at 22.5M (inside the pre-RL interval), $0.0723 \rightarrow 0.0650$ at 56M, $0.0817 \rightarrow 0.0774$ at 155M and $0.0736 \rightarrow 0.0700$ at 508M. The dividend is not bought with the belief.

F THE DIVIDEND ON A LANGUAGE-MODEL LADDER

Pre-registered and run on 2026-09-22 (plan section "LM ladder"). **Why.** The closest analogue of the family in the language-model world is a ladder of models pretrained on one corpus with one recipe, cloned from one fixed teacher and then trained by RL with an exact reward; Pythia (Biderman et al., 2023) (14M, 31M, 70M, 160M, 410M; the Pile, the same data order at every size) is that ladder; the 14M and 31M checkpoints postdate the Pythia paper, and the five models are the Hugging Face revisions cf967c0a (14M), e556ace2 (31M), a39f36b1 (70M), 50f5173d (160M) and 9879c9b5 (410M) (EleutherAI, 2023). **Design.** Task: sort five to eight distinct integers in $[0, 99]$ ("Sort: 57 3 91 12 45 8 =>" $\rightarrow$ "3 8 12 45 57 91"), exact-match reward. Three teachers, as in the small games: A perfect; B systematic, correct except that it drops the last element whenever the list has eight numbers (held-out accuracy 0.748); C noisy, the correct answer with probability 0.7 and otherwise one random adjacent pair swapped (0.709). Clone: full fine-tuning on 50,000 teacher answers (the same prompts at every size for a given seed), cross-entropy on the answer tokens, three epochs, batch 64, learning rate 10^{-4}. RL stage from the clone, the clone frozen as reference: an on-policy policy gradient with a group-normalised advantage (each prompt's 8 samples standardised by their mean and standard deviation; 32 prompts $\times$ 8 samples per step at temperature 1), exact-match reward, a per-token KL penalty to the clone ($\beta = 0.04$), learning rate 10^{-5}, 200 steps; the same recipe at every size. Accuracy on 1,000 fixed held-out prompts under sampling at temperature 1 (the policy as trained; the dividend is RL minus clone) and under greedy decoding; five seeds per cell (the seed sets the teacher's prompts, the SFT order and the RL sampling); 75 cells of one to five minutes each on one RTX 5090. A 1.4B rung was attempted and did not fit: with the same recipe

(full fine-tuning in fp32 with AdamW and a frozen reference) the process holds 26–28 GiB at the RL step when it reaches its 28.6 GiB cap on a 32 GB card shared with the desktop, where without the cap the driver pages instead of failing; making it fit would have put the top rung on a different recipe, so the ladder stops at 410M.

Reading. As for the small games: a size *fits* the teacher, the *plateau* is the sizes that fit, the dividend is *realised* at a size when a majority of seeds are positive, and the shape over the plateau is *size-invariant* (every median within 25% of the grand median), *grows* (monotone with at most one inversion, last − first above the sum of the end IQRs) or neither. The pre-registered fit criterion (the clone's sampled accuracy within 0.05 of the teacher's) was the wrong instrument and was corrected twice, each time with what had been seen written into the plan before the correction: a sampled clone carries its own per-token residue under the teacher (0.59–0.68 against 0.71 for teacher C) while its greedy answer is the teacher's (0.86–0.94), so the criterion became the clone's greedy accuracy at or above the teacher's; a clone of teacher B has the teacher's flawed answer as its mode and lands a few points under it (0.73–0.74 against 0.75), so the original tolerance was restored, greedy accuracy within 0.05 of the teacher's or above. Table 5 shows all three plateaus; the shape rule never changed.

Result (Figure 5, Table 5). Teacher C: the dividend is $+0.328$, $+0.338$, $+0.325$, $+0.343$ and $+0.310$ from 14M to 410M (medians of five seeds, every seed positive, every median within 6% of the grand median), size-invariant on either corrected plateau; under the pre-registered criterion only 410M fits and the shape rule has too few sizes. The prediction carried over from the small games holds. Teacher A: $+0.036$, $+0.037$, $+0.019$, $+0.022$, $+0.018$, the clone's own sampling residue (the clone samples the correct answer 0.95–0.98 of the time, the RL stage 0.99–1.00), smaller at the larger sizes; the pre-registered "nil" threshold of 0.02 fails by the letter at three of five sizes, and the paper says "only the clone's own sampling residue" rather than "none". Teacher B: $+0.029$, $+0.254$, $+0.017$, $+0.013$, $+0.013$, with 2, 3, 0, 1 and 2 of five seeds discovering the missing element within 200 steps (a dividend of $+0.25$ to $+0.29$, plus one partial fix of $+0.12$ at 410M counted among its two) and the rest keeping the flaw; shape neither; the prediction that this dividend grows with capacity fails. What the ladder shows is the mahjong family's pattern in a second setting: a noisy teacher's headroom is taken at the same rate by every size, and a deterministic flaw is fixed or not fixed run by run, not size by size. What it does not show: the task is synthetic, the students are pretrained on natural text, the recipe is one and the budget short; it is a pilot beside the two games, not a language-model result in its own right.

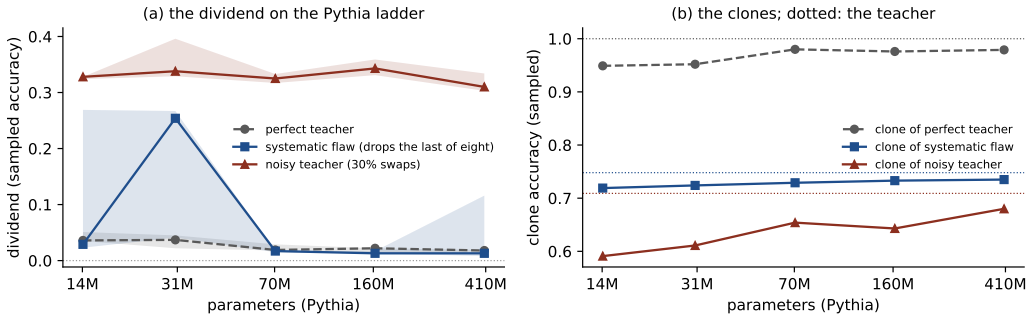


Figure 5: The LM-ladder pilot. (a) The dividend (RL minus clone accuracy under sampling; median over five seeds, interquartile band) against parameters for the three teachers. (b) The clones' sampled accuracy, with each teacher's accuracy dotted.

teacher	size	clone (s / g)	RL (s)	dividend, median [IQR]	seeds > 0.1	fits (pre / c1 / c2)
A perfect	14m	0.949 / 0.973	0.991	$+0.036$ $[+0.035, +0.051]$	0 / 5	– / – / ✓
	31m	0.952 / 0.975	0.994	$+0.037$ $[+0.022, +0.045]$	0 / 5	✓ / – / ✓
	70m	0.980 / 0.980	0.998	$+0.019$ $[+0.018, +0.029]$	0 / 5	✓ / – / ✓
	160m	0.976 / 0.985	0.997	$+0.022$ $[+0.014, +0.022]$	0 / 5	✓ / – / ✓
	410m	0.979 / 0.986	0.998	$+0.018$ $[+0.016, +0.019]$	0 / 5	✓ / – / ✓
B systematic flaw	14m	0.719 / 0.736	0.745	$+0.029$ $[+0.023, +0.269]$	2 / 5	✓ / – / ✓
	31m	0.724 / 0.730	0.993	$+0.254$ $[+0.041, +0.267]$	3 / 5	✓ / – / ✓
	70m	0.729 / 0.738	0.744	$+0.017$ $[+0.015, +0.017]$	0 / 5	✓ / – / ✓
	160m	0.733 / 0.736	0.746	$+0.013$ $[+0.013, +0.017]$	1 / 5	✓ / – / ✓
	410m	0.735 / 0.741	0.748	$+0.013$ $[+0.008, +0.116]$	2 / 5	✓ / – / ✓
C noisy	14m	0.591 / 0.863	0.919	$+0.328$ $[+0.324, +0.328]$	5 / 5	– / ✓ / ✓
	31m	0.611 / 0.886	0.944	$+0.338$ $[+0.329, +0.396]$	5 / 5	– / ✓ / ✓
	70m	0.654 / 0.898	0.983	$+0.325$ $[+0.317, +0.333]$	5 / 5	– / ✓ / ✓
	160m	0.643 / 0.932	0.985	$+0.343$ $[+0.331, +0.359]$	5 / 5	– / ✓ / ✓
	410m	0.680 / 0.936	0.988	$+0.310$ $[+0.303, +0.334]$	5 / 5	✓ / ✓ / ✓

Table 5: The LM-ladder pilot, generated from the report: per teacher and size, the clone's accuracy (sampled / greedy, medians over seeds), the RL stage's sampled accuracy, the dividend (median and IQR over seeds), the number of seeds whose dividend exceeds 0.1, and whether the size fits the teacher under the pre-registered criterion (pre) and the two corrections (c1, c2) of Appendix F.