COOPSHIFT: POST-TRAINING AMPLIFIES THE MEMORY CURSE IN LLM COOPERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

How does post-training change cooperation during repeated interaction? Within each of seven released model lineages, we compare pretrained (Base), supervised fine-tuned (SFT), and subsequent preference- or reward-optimized (Aligned) checkpoints across seven social dilemmas. Copies of each checkpoint play one another for 200 rounds, with visible history limited to the last two or forty rounds. The memory curse refers to lower cooperation when agents can see more past rounds. Our central finding is that post-training amplifies this decline. Averaged across the six binary-action games, expanding visible history reduces cooperation at every SFT and Aligned checkpoint. At each post-trained stage, these declines exceed Base's in six of seven lineages. Most of this average amplification is already present at SFT. With long history, both post-trained stages also cooperate less than Base on average in each binary-action game. In the symmetric binary games, their fitted cooperation rates fall more steeply on average than Base's as per-round cooperation costs rise, benefits fall, or noncooperation becomes more rewarding. The difference in cooperation rates between Aligned and SFT also changes from one-shot decisions to repeated play. When given the same past records, their choices depend on whether the instructions describe a final round or possible future interaction. Post-training evaluation should therefore go beyond isolated choices to assess whether cooperation is sustained in repeated interaction and how it depends on payoffs and visible history.

1 INTRODUCTION

Post-training adapts models to follow instructions and satisfy evaluated preferences (Ouyang et al., 2022; Rafailov et al., 2023). Repeated interaction adds a challenge: sustaining cooperation despite incentives to deviate (Piatti et al., 2024). In a public-goods game (PGG), for example, players keep their endowments or contribute them to a pool. The pool is multiplied by a factor greater than one and shared equally among all players, including noncontributors. In our PGG, mutual contribution maximizes group payoff and gives everyone more than mutual retention. Yet withholding one's contribution while others contribute earns that player even more this round, so mutual contribution is not a one-round Nash equilibrium. Repetition changes the trade-off: the immediate gain from defection can be outweighed by losing future mutual cooperation (Fudenberg & Maskin, 1986). We ask: *how does cooperation differ across training stages, and how do these differences change when agents can see more of their interaction history?*

Past interaction matters because agents can use it to decide whether to keep cooperating, withdraw, or try to restore cooperation. A longer record gives them more evidence about their partners; its effect on cooperation depends on how that evidence changes their choices. Human experiments separate private contribution costs from public returns (Goeree et al., 2002), reveal payoff-dependent conditional cooperation (Baader et al., 2026), and link declining contributions to imperfect conditional cooperation (Fischbacher & Gächter, 2010). In human experiments, intermediate visible histories can also support more cooperation than longer records (Ma et al., 2021).

Language-model cooperation depends on objectives, partners, and repeated interaction (Phelps & Russell, 2025; Akata et al., 2025). Liu et al. (2026) document a *memory curse*: agents can cooperate less when their prompts contain more past rounds. Here, *memory* means the interaction record visible at each decision. Within each model family, or *lineage*, we compare pretrained Base, demonstration-

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.

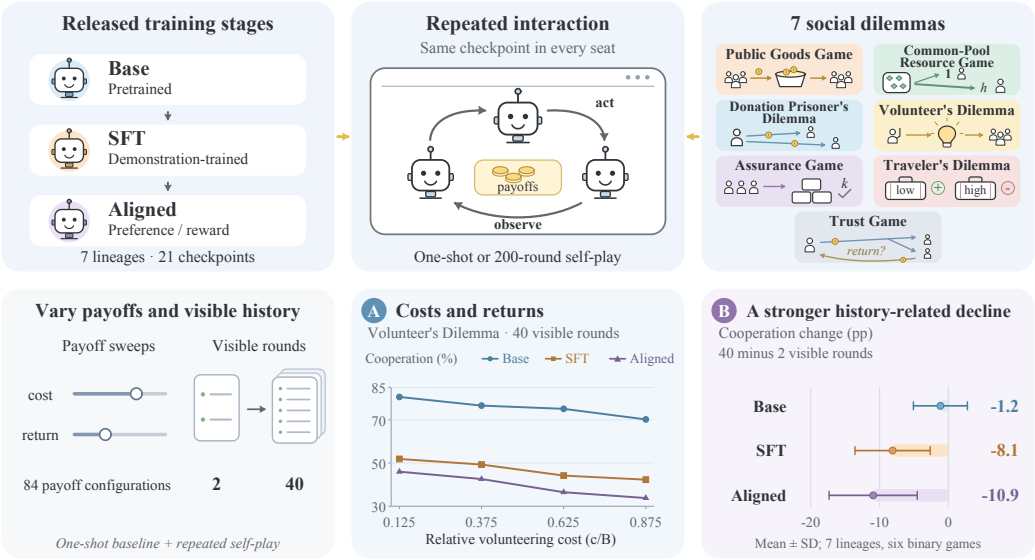

Figure 1: **CoopShift: cooperation across released post-training stages.** Seven lineages (21 checkpoints) play seven social dilemmas across 84 payoff configurations. One-shot decisions complement 200-round self-play. **A:** Volunteer's Dilemma with a forty-round history limit (HL40): mean cooperation across seven lineages versus relative volunteering cost c/B, where c is cost and B each player's benefit. The y axis spans 30–85%; Table 20 shows each lineage separately. **B:** Forty-minus-two-round history responses, in percentage points (pp): mean ± standard deviation (SD) across seven lineage summaries, each averaging the six binary games (Table 3).

trained SFT, and a subsequent preference- or reward-optimized Aligned checkpoint. We ask whether expanding visible history produces a larger cooperation decline at the post-trained checkpoints than at Base. This larger short-to-long-history decline is what we call *amplification* of the memory curse.

We evaluate 21 checkpoints from seven lineages in seven social dilemmas: PGG, a common-pool resource game (CPR), Donation Prisoner's Dilemma (Donation-PD), the Volunteer's Dilemma (Volunteer), Assurance, Traveler's Dilemma (TD), and the Trust game (Table 1; Appendix C). One-shot prompts present a single decision without past interaction or future-round statements. In repeated play, copies of one checkpoint interact for 200 recorded rounds, seeing up to the two or forty most recent rounds. Separate tests give different checkpoints the same past-interaction records, either constructed by the researchers or saved from earlier games, and compare their current choices.

For each binary game, the cooperation rate records how often its designated cooperative action is chosen. At a fixed PGG or Donation-PD payoff setting, this rate also determines the group's average earned points. We use *collective welfare* to mean the sum of all players' earned points.

Our central finding is that **post-training amplifies the memory curse (Section 4).** After averaging the six binary games equally within each lineage, increasing visible history from two to forty rounds lowers cooperation at every SFT and Aligned checkpoint, by about 8 and 11 percentage points (pp) on average, respectively. Base's mean decline is about 1 pp, and its cooperation increases in five of seven lineages. At each post-trained stage, the decline exceeds Base's in six of seven lineages, and most of the average amplification is already present at SFT. We extend prior memory evidence by showing where the response to longer histories becomes stronger along released training sequences.

With forty visible rounds, SFT and Aligned cooperate less than Base on average in each binary game. **Their cooperation rates also fall more than Base's, on average, as per-round cooperation costs rise, benefits fall, or noncooperation becomes more rewarding (Section 3).** This comparison uses fitted trends across each game's tested payoff values and holds in the symmetric binary games. TD has a different outcome: both stages choose higher mean claims than Base. Lower contributions translate into lower mean PGG and Donation-PD payoffs; Section 7 relates these group outcomes to individual action quality. Differences in cooperation between stages therefore depend on both payoffs

and visible history. Section 5 compares one-shot with repeated play, then gives checkpoints the same past records but different future instructions.

2 Experimental Design

We compare 21 checkpoints from seven open-weight lineages. Aligned objectives correspond to direct preference optimization (DPO), reinforcement learning (RL) from human feedback (RLHF), or group relative policy optimization (GRPO). In Base / SFT / Aligned order, the display names are: OLMo-3-7B (Base / Instruct-SFT / Instruct-DPO); SmolLM2-1.7B (Base / SFT-only / Instruct); Mistral/Zephyr-7B (Mistral-v0.1 / Mistral-SFT-Beta / Zephyr-Beta); Llama/Tülu-3-8B (Llama-3.1 / Tülu-3-SFT / Tülu-3-DPO); InternLM2-7B (Base / Chat-SFT / Chat); DeepSeekMath-7B (Base / Instruct / RL); and Nemotron-Cascade-8B (Qwen3-Base / Cascade-SFT / Cascade-RLHF). We abbreviate stages as B, S, A. Appendix B, Table 7, gives exact Hugging Face repository identifiers and links, training objectives, and checkpoint ancestry.

A *payoff configuration* fixes a game's player count and all its payoff parameters. We evaluate every checkpoint on the same 84 configurations: PGG (8), CPR (16), Donation-PD (12), Volunteer (12), Assurance (8), TD (16), and Trust (12). Each is used in one-shot, HL2, and HL40 evaluations. In the first five games, players have the same two action options and payoff rule; TD offers multiple claims; Trust has an investor and two trustees. Table 1 explains the parameters; Appendices C and Q give payoff matrices, functions, and complete results.

Table 1: What the games measure. The main parameter changes a cost, return, or coordination reward, depending on the game. Appendix C gives the full payoff rules and all other parameter values.

Game	Decision and tension	Meaning of the raw main parameter
PGG	Contribute a unit to a shared pool or retain it.	The marginal per-capita return (MPCR), $\alpha = r/n$, is each player's return per contributed unit, where r is the pool multiplier and n the group size. Higher α lowers private cost $1 - \alpha$.
CPR	Restrain extraction or extract more; each unit earns less as total extraction rises.	Low extraction is 1 and high extraction is $h > 1$. Higher θ increases high extraction's private advantage at fixed h; no resource stock carries across rounds.
Donation-PD	Pay c to give b to each other player, or give nothing.	c is the donor's cost and b the benefit to each of two others. Higher $R = 2b/c$, at fixed c, raises external returns; private cost stays c.
Volunteer	Supply a shared benefit or rely on another volunteer; one suffices.	Each player receives B if anyone volunteers; every volunteer pays c, so extra volunteers duplicate costs. Higher $\rho = c/B$, at fixed B, increases volunteering cost.
Assurance	Join a threshold project or take a sure outside payoff.	A successful contributor earns a fraction λ more than the sure outside payoff. Success requires at least k contributors, including itself; unsuccessful contributors receive zero. Higher λ raises the reward for successful coordination.
TD	Claim an amount; the lower claimant gets a bonus, the higher a penalty.	K is the number of legal claim levels. Higher τ, at fixed K, raises bonus/penalty and one-step undercutting gain. Outcome: claim normalized to $[0, 1]$, not binary cooperation.
Trust	Choose whether to invest; two trustees state simultaneous, contingent return intentions.	Investment gives each trustee $5m$, where m is the multiplier. A returning trustee sends back $5(1 + s)$; higher s raises the investor's return and the trustee's cost. Investment and return intentions are separate outcomes.

Whose cooperation is measured? Within each run, we average players and then rounds. We then average runs and payoff configurations. Thus Trust counts each of its three players equally: the investor has weight $1/3$ and the two trustees together $2/3$ (Appendix I). The separate shared-record tests score only the player asked to decide.

One-shot and repeated decisions. One-shot prompts omit history and future-round statements. Repeated actions are simultaneous. Following Tewolde et al. (2026) and Liu et al. (2026), prompts describe probabilistic continuation and a long-run payoff objective. Our prompts state a 99% probability of another round; recording uses a fixed 200 rounds without announcing this endpoint. The history limit (HL) caps visible completed rounds at two (HL2) or forty (HL40); early in play,

agents see all available rounds up to that cap. Each configuration uses five trajectory seeds or 20 one-shot draws. Section 5.2 defines the separate final-round and continuing probes.

Response formats. In the main one-shot and repeated-game evaluation, the model gives a rationale and then selects a legal action label (prompt template: Appendix M.1; sampling: Appendix M). The shared-record tests use the interfaces documented in Appendices K and M.

What we measure. The cooperation rate is the fraction of decisions taking the game's designated cooperative action. We report rates as percentages and changes as percentage points (pp): a fall from 70% to 60% is a change of -10 pp. The six binary-action games, including Trust, form our *cooperation panel*; overall cooperation and history summaries weight these games equally. TD instead measures the chosen claim, scaled from zero for the lowest legal claim to one for the highest, and is reported separately.

How to read the summaries. Each lineage contributes one estimate to an overall mean. We first average runs and payoff settings within that lineage; summaries of the six binary games then weight those games equally. To compare stages, we compute their difference within each lineage before averaging the seven differences. Tables report the mean and, where shown, the SD across these seven estimates. The SD describes how much the models differ; a count such as 6/7 says how many lineages show the stated direction. Intervals give a range for the mean, based on either variation across lineages or repeated sampling with the same models, as specified in each table. Table 2 reports means; Appendix H shows each lineage separately.

What changes when we compare stages? In self-play, a checkpoint plays against copies of itself. Replacing Base with SFT therefore changes both a player and its partners; their choices also change the history seen in later rounds. Our main comparisons measure the group behavior each checkpoint produces under the same game rules and history limit. The tests in Section 5.2 ask a complementary question: how do checkpoints choose when given the same past records?

3 Cooperation Rates and Payoff Responses

At HL40, Table 2 reports mean cooperation at four values of each game's main payoff parameter, labeled D_1–D_4; the parameter row gives the game-specific values. From D_1 to D_4, cooperation costs rise, benefits fall, or noncooperation becomes more rewarding. For example, lower PGG marginal per-capita return (MPCR)—each player's return per unit contributed—raises the private contribution cost.

Cooperation levels. Compare stages vertically within the same column. In every symmetric binary game, SFT and Aligned are below Base at all four displayed payoff values: they cooperate less under the same incentives. Averaging seven lineages at each complete binary-game configuration, SFT is below Base in 66/68 and Aligned in 64/68 (Appendix D). Trust shows the same game-level ordering; TD instead has higher post-trained mean normalized claims.

Responses to changing payoffs. Read across a row, then compare its Drop with the other stages'. For each lineage and stage, a straight-line fit uses all four payoff values at their actual numerical spacing. The *fitted drop* is the fitted rate at D_1 minus that at D_4, in pp. Positive values indicate a downward fitted trend; near-zero values indicate little net fitted change. Individual steps can still rise (Appendix A.1).

In each symmetric binary game, both post-trained mean drops are positive and larger than Base's: cooperation falls more as per-round costs rise, benefits fall, or noncooperation becomes more rewarding (from D_1 to D_4). To see the stage gaps in Table 2, subtract Base from SFT or Aligned within each column (differences listed in Table 16). In PGG, the SFT-minus-Base gap changes from -16.6 pp at D_1 to -23.5 pp at D_4: SFT falls further below Base. Comparing the Drop entries summarizes how these gaps change using all four points. Volunteer has a large cooperation gap but nearly parallel SFT and Base trends. Appendix H shows each lineage separately.

Do less cooperative groups also earn less? PGG and Donation-PD let us answer this directly: at a fixed payoff setting, every contribution adds a fixed positive amount to the group's total, so the contribution rate determines average points earned per player. Other games require more information about joint actions; for example, Assurance payoffs depend on whether enough players contribute to complete the project. We therefore convert rates to returns in these two games before averaging

Table 2: **Cooperation across four payoff values at HL40.** Entries are means across seven lineages (%; normalized claims for TD), after averaging seeds and secondary parameters. The final column is the fitted decline across all four points (pp); the numerical parameter spacing is retained. Table 16 adds mean paired stage differences; Tables 17–22 show every lineage with seed SDs.

Game / stage	D_1	D_2	D_3	D_4	Drop
PGG: MPCR	0.8	0.6	0.5	0.4	–
Base (B)	59.8	58.6	57.1	57.2	+2.9
SFT (S)	43.2	38.6	35.4	33.7	+9.8
Aligned (A)	42.2	38.7	35.6	31.4	+10.5
CPR: θ	0.2	0.4	0.6	0.8	–
Base (B)	68.0	63.7	62.3	65.6	+2.6
SFT (S)	31.4	33.2	30.5	27.1	+4.6
Aligned (A)	23.6	26.2	24.6	19.5	+4.2
Donation-PD: R	2.4	2	1.4	1.1	–
Base (B)	54.3	52.9	53.7	54.3	−0.2
SFT (S)	15.9	12.6	13.6	13.4	+1.8
Aligned (A)	11.7	8.9	9.3	8.7	+2.3

Game / stage	D_1	D_2	D_3	D_4	Drop
Volunteer: ρ	0.125	0.375	0.625	0.875	–
Base (B)	80.6	76.6	75.1	70.2	+9.8
SFT (S)	51.9	49.3	44.2	42.3	+10.2
Aligned (A)	46.0	42.6	36.5	33.8	+12.8
Assurance: λ	0.4	0.2	0.1	0.05	–
Base (B)	87.2	85.7	86.8	87.3	+0.0
SFT (S)	80.8	78.3	71.5	75.0	+7.7
Aligned (A)	85.2	80.0	70.4	76.4	+12.0
TD: τ	0.025	0.05	0.1	0.2	–
Base (B)	39.6	39.8	40.7	37.9	+1.7
SFT (S)	67.2	66.9	67.0	67.4	−0.3
Aligned (A)	62.6	62.1	61.4	60.4	+2.2

settings. Both post-trained stages earn fewer points than Base on average at HL40, with decreases more consistent across lineages in Donation-PD (Appendix R.1).

4 Post-Training Amplifies the Memory Curse

Does the difference in cooperation between stages also change when more past interaction is visible? We compare separate 200-round HL2 and HL40 runs of the same checkpoints at the same complete payoff configurations. Following Liu et al. (2026), the *memory curse* means lower cooperation with a longer visible history. Averaging the six binary games equally within each lineage, mean cooperation falls by about 8 pp for SFT and 11 pp for Aligned, compared with about 1 pp for Base.

Both post-trained stages decline in all seven lineage summaries and decline more than Base in six; Base cooperation increases in five. Most of the average amplification is already present at SFT: its decline exceeds Base's by 6.93 pp, compared with a further 2.82 pp from SFT to Aligned. Table 3 shows that both post-trained stages' additional declines relative to Base have intervals below zero. The smaller difference between SFT and Aligned has an interval spanning zero.

Table 3: **The memory curse and its amplification after post-training.** The first three rows report HL40 minus HL2; the last three compare these changes between stages. Entries give mean ± SD across seven lineages, in pp, after averaging the six binary games equally within each lineage. Stage differences are calculated within lineage. The exploratory 95% t intervals use these seven estimates; the last column counts negative estimates.

History change or paired difference	Mean ± SD	95% interval	Negative
Base: HL40 minus HL2	-1.17 ± 3.92	$[-4.79, +2.45]$	2/7
SFT: HL40 minus HL2	-8.11 ± 5.46	$[-13.16, -3.05]$	7/7
Aligned: HL40 minus HL2	-10.93 ± 6.40	$[-16.85, -5.01]$	7/7
SFT minus Base history change	-6.93 ± 7.30	$[-13.68, -0.19]$	6/7
Aligned minus Base history change	-9.76 ± 7.47	$[-16.67, -2.85]$	6/7
Aligned minus SFT history change	-2.82 ± 4.37	$[-6.87, +1.22]$	6/7

Let $p_s(L)$ be the expected cooperation rate, as a fraction, at stage s with history limit L. The history change H_s compares HL40 with HL2; $K_{s/t}$ compares that change between stages:

$$H_s = p_s(40) - p_s(2), \qquad K_{s/t} = H_s - H_t. \tag{1}$$

A negative H_s means that showing more history lowers cooperation at that stage—the memory curse. $K_{S/B}$ compares SFT with Base, $K_{A/B}$ compares Aligned with Base, and $K_{A/S}$ compares Aligned

with SFT. We multiply these rate differences by 100 when reporting pp. For example, $K_{S/B} = -0.10$, displayed as -10 pp, means that showing more history reduces SFT cooperation by ten percentage points more than Base cooperation.

Figure 2 separates the games and payoff values. A point left of zero means that cooperation falls more, or rises less, at the first-named stage than at the second. Black diamonds show game averages; colored points show individual lineage and payoff results. SFT's mean history change is more negative than Base's in PGG, CPR, Donation-PD, and Volunteer; Assurance has the opposite ordering. Aligned declines more than Base and SFT in all five symmetric binary-game means; in Trust, SFT declines more than Aligned. Longer history also lowers post-trained models' mean returns in PGG and Donation-PD. Appendix O gives each lineage's result, E compares payoff grids, and G shows every complete configuration.

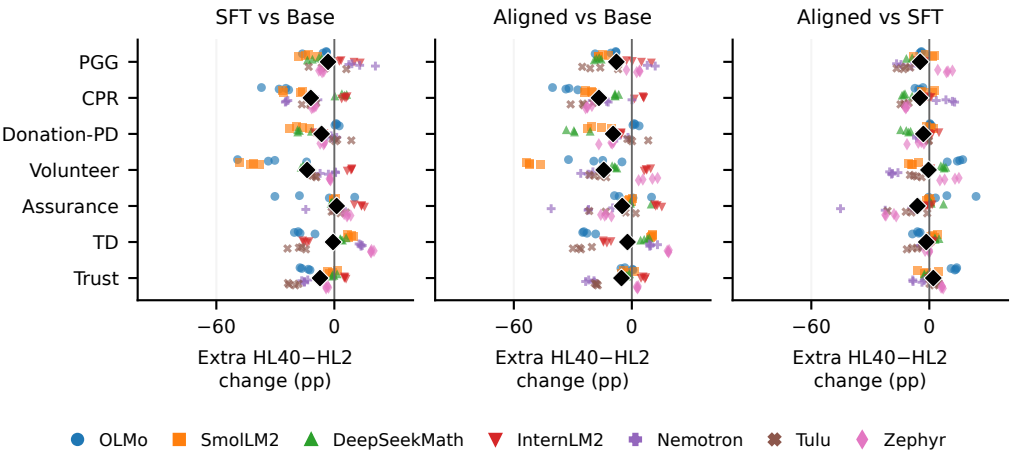

Figure 2: **Which stage loses more cooperation when shown a longer history?** Panels compare SFT vs Base ($K_{S/B}$), Aligned vs Base ($K_{A/B}$), and Aligned vs SFT ($K_{A/S}$). Left of zero means a more negative change from HL2 to HL40 at the first-named stage: a larger decrease or smaller increase. TD uses normalized claims rather than binary cooperation. Each game/panel has 28 colored points: seven lineages $\times$ four primary payoff values, averaging five seeds and secondary parameters. Black diamonds average points equally within and across lineages. All three panels use the same percentage-point scale.

A testable explanation. A model that reduces cooperation after a partner defects changes what that partner sees next. If neither readily resumes cooperating, their responses can reinforce one another over time. Post-training may change this balance between withdrawal and recovery. Human PGG experiments show a related pattern: contributing less than others helps drive declining cooperation (Fischbacher & Gächter, 2010). Our separate test of older defection records finds game-specific responses, strongest after post-training in CPR (Appendix J). Tracking recovery after a temporary defection would test the proposed feedback process directly.

Intermediate stages. Nemotron-Cascade adds instruction-following, mathematics, code, and software-engineering RL after RLHF, forming seven checkpoints. Each post-Base checkpoint has a more negative mean history change than Base across the six binary games. Instruction-following RL nevertheless raises cooperation relative to RLHF in one-shot, HL2, and HL40 evaluations. Cooperation can therefore recover at a later training stage while the stronger memory curse remains (Appendix T).

5 Cooperation across Interaction Settings

The history result raises a further question: does a one-shot cooperation advantage persist in actual interaction? We focus on Aligned versus SFT to examine training after SFT, then compare their choices given the same past records.

5.1 ALIGNED VERSUS SFT IN ONE-SHOT AND REPEATED PLAY

Table 4 asks whether Aligned cooperates more than SFT. Positive entries mean that it does. The first two columns give this difference in one-shot decisions and repeated play; the third shows how much it changes. We calculate changes within each lineage before averaging them. A negative change means that moving to repeated play shifts the difference toward SFT (Appendix A.1).

The difference decreases in all five symmetric binary-game means. In PGG, CPR, and Volunteer, Aligned cooperates more than SFT in one-shot decisions but less in repeated play. In Donation-PD, it cooperates less in both; Assurance changes little. Volunteer has the clearest pattern: the difference falls by about 10 pp, with all seven lineages showing a decrease. Other games have mixed lineage directions. TD's mean claim difference also decreases, whereas Trust's mean intention difference increases. A one-shot cooperation advantage can thus disappear when checkpoints actually interact.

Table 4: **Does Aligned cooperate more than SFT in one-shot decisions and repeated play?** The first two columns report Aligned minus SFT; positive values mean more cooperation at Aligned. Change is repeated minus one-shot, computed within each lineage before averaging; negative values mean a smaller Aligned-minus-SFT difference in repeated play. These columns give mean $\pm$ SD across seven model lineages, in pp. SD describes between-model differences in effect magnitude or direction, which can be large. We focus on the mean differences; the last column counts lineages with a negative change. TD measures normalized claims; Trust averages roles.

Game	One-shot Aligned − SFT	Repeated (HL40) Aligned − SFT	Change Repeated − one-shot	Lineages with a decrease
PGG	$+8.62 \pm 6.78$	-0.77 ± 20.44	-9.39 ± 20.00	5/7
CPR	$+3.47 \pm 22.35$	-7.08 ± 14.96	-10.55 ± 19.14	5/7
Donation-PD	-1.71 ± 7.11	-4.21 ± 5.51	-2.51 ± 9.25	4/7
Volunteer	$+2.66 \pm 14.47$	-7.25 ± 11.63	-9.91 ± 5.18	7/7
Assurance	$+1.93 \pm 13.67$	$+1.59 \pm 8.14$	-0.35 ± 8.96	3/7
TD	$+2.48 \pm 9.71$	-5.50 ± 8.82	-7.98 ± 12.26	5/7
Trust	-3.91 ± 8.47	$+0.21 \pm 12.99$	$+4.12 \pm 9.03$	2/7

Interpretation. In one-shot decisions, a model never encounters its partners' responses. Repeated play adds this feedback: its choices become part of the history used for later decisions. Training recipes also differ across lineages. SFT demonstrations already teach desired behavior, and later training uses different preferences and rewards (Appendix B). These differences may help explain why the Aligned–SFT comparison varies across models.

5.2 INSTRUCTIONS ABOUT FUTURE ROUNDS HAVE GAME-SPECIFIC EFFECTS

5.2.1 FINAL-ROUND AND CONTINUING DECISIONS

Self-play gives different checkpoints different histories. We now supply the same scripted records and ask for one current decision under different future instructions across all seven games. FINAL says that the interaction ends after the current round. CONTINUE says that there is a 99% probability of another round, that actions will be visible in future rounds, and that the goal is to maximize long-run points. We keep the payoff, checkpoint, scripted past records, and current round number (41) the same between the two prompts. The comparison measures the effect of changing these future-interaction instructions together.

We supply either no past records or 2, 10, or 40 scripted rounds, using seven behavior patterns at each nonzero length. This gives $1 + 3 \times 7 = 22$ conditions, averaged in Table 5 (Appendix M). These tests measure one decision under each prompt; the earlier repeated-game results instead measure 200 rounds of actual interaction.

In Table 5, the effect of continuing instructions is the CONTINUE rate minus the FINAL rate, denoted Δ and reported in pp. Continuing instructions increase CPR restraint by about 15 pp on average and lower TD claims by about 13 pp. These mean directions hold at every combination of training stage and history length (Appendix L). In CPR, the increase is larger at SFT than Base and larger again at Aligned. In PGG, the small average effect combines an increase at SFT with a decrease at Aligned.

Table 5: **How instructions about future rounds change the current decision.** FINAL says this is the last round; CONTINUE states a 99% chance of another. Cooperation is in percent; TD claims are rescaled from 0 (lowest allowed) to 100 (highest). Δ is CONTINUE minus FINAL, in pp. Entries are mean $\pm$ SD across seven model lineages, after averaging payoff settings and 22 history conditions within lineage. SD captures differences between models, which can be large; we focus on the means. The first three numeric columns also average stages; the last three keep stages separate. Trust equally weights the investor and a representative trustee.

Game	FINAL	CONTINUE	Mean Δ	Δ Base	Δ SFT	Δ Aligned
PGG	73.7 ± 14.3	73.2 ± 14.4	-0.5 ± 4.0	-0.1 ± 3.4	$+2.5 \pm 6.4$	-3.9 ± 6.8
CPR	43.9 ± 18.4	59.0 ± 15.2	$+15.1 \pm 9.7$	$+6.2 \pm 4.0$	$+14.0 \pm 10.4$	$+25.1 \pm 20.6$
Donation-PD	61.0 ± 21.8	64.0 ± 14.5	$+3.0 \pm 9.9$	-0.8 ± 7.2	$+5.5 \pm 12.7$	$+4.4 \pm 12.2$
Volunteer	78.3 ± 13.7	81.3 ± 10.5	$+3.0 \pm 7.5$	$+3.6 \pm 5.1$	$+2.8 \pm 10.8$	$+2.8 \pm 11.0$
Assurance	72.9 ± 18.0	75.1 ± 12.2	$+2.2 \pm 10.8$	$+1.0 \pm 6.3$	$+3.8 \pm 16.6$	$+1.6 \pm 12.9$
TD	53.2 ± 14.6	40.7 ± 9.8	-12.6 ± 8.2	-8.4 ± 9.5	-11.4 ± 7.2	-17.9 ± 14.3
Trust	73.9 ± 11.9	75.4 ± 9.5	$+1.5 \pm 5.9$	-0.2 ± 1.8	$+3.2 \pm 9.3$	$+1.5 \pm 8.0$

Thus the same instructions about future interaction can raise or lower cooperation, depending on the game and checkpoint.

Human repeated-PD experiments find higher cooperation at greater actual continuation probabilities after participants gain experience (Dal Bó & Fréchette, 2011). Our test asks for one choice after telling the model either that this is the final round or that there is a 99% chance of another round. These continuing instructions raise CPR cooperation, lower TD claims, and change PGG cooperation little on average (Table 5; Appendix L).

5.2.2 PAST RECORDS AND INSTRUCTIONS ABOUT FUTURE INTERACTION

Does showing past records have a different effect under final-round and continuing instructions? We give SFT and Aligned fresh PGG decisions at four payoff settings under four combinations: records shown or hidden, each with FINAL or CONTINUE. REAL shows saved actions, earned points, and summaries; MASKED removes them but keeps rules, role, round number, and future instructions. Removing records also shortens the prompt. Table 6 subtracts hidden-record from shown-record cooperation: negative means less cooperation when records are shown (Appendix K).

Table 6: **How showing past records changes PGG cooperation.** REAL shows the records; MASKED hides them. The first two numeric columns give rates (%); the last subtracts the second from the first to give the effect of showing records (pp). Negative values mean less cooperation with records visible. Entries are mean $\pm$ SD across seven lineages. All four conditions use the same payoff and record weights (Appendix K).

Future instruction	Stage	REAL	MASKED	REAL$-$MASKED
CONTINUE	SFT (S)	41.9 ± 31.7	58.2 ± 27.6	-16.3 ± 12.3
	Aligned (A)	41.4 ± 35.3	63.3 ± 29.9	-22.0 ± 12.9
FINAL	SFT (S)	35.8 ± 22.2	51.5 ± 22.0	-15.7 ± 21.9
	Aligned (A)	34.6 ± 26.4	46.6 ± 23.8	-12.0 ± 32.0

Showing past records lowers mean contribution at both stages. Under CONTINUE, the mean reduction is larger for Aligned than for SFT; under FINAL, it is larger for SFT. The ordering of these average history effects therefore reverses when the same decision is described as final rather than potentially continuing.

Subtracting SFT's record effect from Aligned's gives one difference for each frame. This difference falls by 9.42 pp on average when FINAL becomes CONTINUE, with the same direction in five of seven lineages. Resampling lineages gives a wide 97.5% interval, $[-24.29, +2.45]$ pp, which includes zero: model choice matters for this comparison. Appendix K also reports intervals for repeated sampling from the same models and archived records; Appendix P gives their calculation and simulation checks.

Why the final-round result is informative. In final-round PGG, withholding one's contribution always earns $1 - r/n > 0$ more points this round, holding the others' actions fixed. Past records still change the models' choices. One possibility is that models respond to past behavior even when told that there is no future interaction to influence.

6 Related Work

We summarize the closest work here; Appendix S provides the detailed related-work review.

Cooperation and memory in language agents. Prior studies examine how language agents respond to social objectives, partners, and repeated interaction (Phelps & Russell, 2025; Fontana et al., 2025; Akata et al., 2025; Piatti et al., 2024). This literature evaluates both cooperation-sustaining mechanisms, such as repetition, reputation, mediation, and contracting in CoopEval (Tewolde et al., 2026), and history effects, including the memory curse (Liu et al., 2026). CoopShift asks whether the cooperation change from showing more history differs across Base, SFT, and Aligned checkpoints within a released lineage. We complement this comparison with probes that give SFT and Aligned the same past records and with the payoffs earned during self-play. These measure, respectively, choices under common historical inputs and outcomes generated by interacting groups.

Post-training and human behavior. Instruction and preference optimization train models to produce desired responses (Ouyang et al., 2022; Rafailov et al., 2023). We ask how released checkpoints behave when their decisions affect other players. Human experiments provide comparisons for specific responses to costs, benefits, and past interaction; Appendix S explains where their designs match or differ from ours.

7 Discussion and Conclusion

Differences in cooperation between checkpoints depend on both costs and benefits and the history agents can see. Most of the average memory-curse amplification appears between Base and SFT, making demonstration training central to understanding it. Why longer histories lead post-trained models to cooperate less remains open; Section 4 proposes testing withdrawal and recovery over successive rounds. Nemotron's later rise in cooperation shows that a higher cooperation rate can coexist with a larger history-related decline (Appendix T).

Good individual choices and successful cooperation. Against partners with known rules, SFT and Aligned give up fewer expected points through their current choices than Base on average, in PGG, Donation-PD, and Assurance at both history lengths (Appendix U). We score that action assuming the best choices for the remaining 39 decisions. Their PGG and Donation-PD self-play groups nevertheless earn less on average. The distinction is between choosing well against given partners and creating value together: contributions generate the surplus that makes mutual cooperation better for everyone than mutual withholding.

Interpreting Base behavior. Base follows the same procedure of giving a rationale and then choosing an action. Its small average history effect combines different responses: InternLM2 Base and Nemotron's Qwen3 Base decline by about 7.6 and 6.0 pp (Appendix O). Some Base checkpoints therefore respond to history. Swapping action labels and testing payoff comprehension would help determine whether their choices follow game incentives or answer preferences.

Next tests. Holding partner behavior fixed while introducing a temporary defection and then restoring cooperation would distinguish a stronger immediate response from slower recovery. Comparing Base, SFT, and Aligned under this same sequence would locate the training step where either response changes, alongside its effect on earned returns.

Conclusion. Post-training amplifies the memory curse in our six-binary-game panel. At HL40, both post-trained stages cooperate less than Base on average in each binary game. In the symmetric binary games, their fitted cooperation rates also fall more than Base's, on average, as per-round cooperation costs rise, benefits fall, or noncooperation becomes more rewarding. Evaluation should track how cooperation changes with payoffs and visible history, what returns it produces, and whether it recovers when partners resume contributing.

Reproducibility Statement

Section 2 outlines the experimental design and comparisons. Appendix B (Table 7) provides the exact Hugging Face repository identifiers and links for all 21 Base, SFT, and Aligned checkpoints, together with their training objectives and ancestry; Appendix T documents the extended seven-stage Nemotron comparison. Appendix M.1 provides the main prompt template, history format, one-shot variant, and rationale-to-action message sequence. Full game rules and payoff grids appear in Appendix C; measurement, aggregation, and uncertainty definitions in Appendix A; and inference settings, data coverage, and retry procedures in Appendices M–N. Appendices K–L and P detail the shared-input probes and their sampling procedures; Appendix U specifies the programmed-partner case study.

AI Use Statement

Generative AI assisted conceptual framing, hypothesis refinement, experimental-design feedback, elementary payoff derivations and the Nash-reversion benchmark in Appendix R, implementation, data cleaning and reformatting, statistical cross-checks, result interpretation, literature synthesis, manuscript drafting, and figure/table preparation. The evaluated agents themselves generate the synthetic game decisions. The authors are responsible for scientific validation and the paper's text, figures, tables, and claims.

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, 2025.

Lisa R Anderson, Robert L Hicks, and Andrew Turscak. Cooperation in public goods experiments with random and finite stopping rules. *Journal of the Economic Science Association*, 10(2): 367–392, 2024.

James Andreoni. Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *The Quarterly Journal of Economics*, 110(1):1–21, 1995.

Malte Baader, Simon Gächter, Kyeongtae Lee, and Martin Sefton. Social preferences and the variability of conditional cooperation. *Economic Theory*, 81(1):341–366, 2026.

Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142, 1995.

Marcel Binz, Elif Akata, Abdullah Almaatouq, Mohammed Alsobay, Oleksii Ariasov, Franziska Brändle, David Broska, Jason W Burton, Nuno Busch, Frederick Callaway, et al. Post-training makes large language models less human-like. *arXiv preprint arXiv:2605.07632*, 2026.

C Monica Capra, Jacob K Goeree, Rosario Gomez, and Charles A Holt. Anomalous behavior in a traveler's dilemma? *American Economic Review*, 89(3):678–690, 1999.

Gary Charness and Matthew Rabin. Understanding social preferences with simple tests. *The quarterly journal of economics*, 117(3):817–869, 2002.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

Pedro Dal Bó and Guillaume R Fréchette. The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review*, 101(1):411–429, 2011.

Urs Fischbacher and Simon Gächter. Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American economic review*, 100(1):541–556, 2010.

Urs Fischbacher, Simon Gächter, and Ernst Fehr. Are people conditionally cooperative? evidence from a public goods experiment. *Economics letters*, 71(3):397–404, 2001.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. Nicer than humans: How do large language models behave in the prisoner's dilemma? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pp. 522–535, 2025.

Drew Fudenberg and Eric Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, 1986.

Drew Fudenberg, David G Rand, and Anna Dreber. Slow to anger and fast to forgive: Cooperation in an uncertain world. *American Economic Review*, 102(2):720–749, 2012.

Jacob K Goeree, Charles A Holt, and Susan K Laury. Private costs and public benefits: unraveling the effects of altruism and noisy behavior. *Journal of public Economics*, 83(2):255–276, 2002.

R Mark Isaac and James M Walker. Group size effects in public goods provision: The voluntary contributions mechanism. *The Quarterly Journal of Economics*, 103(1):179–199, 1988.

Yuxuan Li and Hirokazu Shirado. Spontaneous giving and calculated greed in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5271–5286, 2025.

Yunhao Liang, Yuan Qu, Jingyuan Yang, Shaochong Lin, and Zuo-Jun Max Shen. Everyone contributes! incentivizing strategic cooperation in multi-llm systems via sequential public goods games. *arXiv preprint arXiv:2508.02076*, 2025.

Jiayuan Liu, Tianqin Li, Shiyi Du, Xin Luo, Haoxuan Zeng, Emanuel Tewolde, Tai Sing Lee, Tonghan Wang, Carl Kingsford, and Vincent Conitzer. The memory curse: How expanded recall erodes cooperative intent in llm agents. *arXiv preprint arXiv:2605.08060*, 2026.

Volodymyr Lugovskyy, Daniela Puzzello, Andrea Sorensen, James Walker, and Arlington Williams. An experimental study of finitely and infinitely repeated linear public goods games. *Games and Economic Behavior*, 102:286–302, 2017.

Shuangmei Ma, Boyu Zhang, Shinan Cao, Jun S Liu, and Wen-Xu Wang. Limited memory optimizes cooperation in social dilemma experiments. *Royal Society open science*, 8(8):210653, 2021.

Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9): e2313925121, 2024.

Long Ouyang, Jeffrey 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. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Steve Phelps and Yvan I Russell. The machine psychology of cooperation: can gpt models operationalize prompts for altruism, cooperation, competitiveness, and selfishness in economic games? *Journal of Physics: Complexity*, 6(1):015018, 2025.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37:111715–111759, 2024.

David Guzman Piedrahita, Yongjin Yang, Mrinmaya Sachan, Giorgia Ramponi, Bernhard Schölkopf, and Zhijing Jin. Corrupted by reasoning: Reasoning language models become free-riders in public goods games. *arXiv preprint arXiv:2506.23276*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Jesse Roberts, Kyle Moore, and Douglas Fisher. Do large language models learn human-like strategic preferences? In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pp. 97–108, 2025.

Eilam Shapira, Moshe Tennenholtz, and Roi Reichart. Alignment makes language models normative, not descriptive. *arXiv preprint arXiv:2603.17218*, 2026.

Natalie Struwe, Esther Blanco, and James M Walker. Increasing benefits in one-time public goods does not promote cooperation. *Proceedings of the National Academy of Sciences*, 121(41): e2410326121, 2024.

Emanuel Tewolde, Xiao Zhang, David Guzman Piedrahita, Vincent Conitzer, and Zhijing Jin. Coopeval: Benchmarking cooperation-sustaining mechanisms and llm agents in social dilemmas. *arXiv preprint arXiv:2604.15267*, 2026.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. Sotopia-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12912–12940, 2024.

Haofei Yu, Zhengyang Qi, Yining Zhao, Kolby Nottingham, Keyang Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. Sotopia-rl: Reward design for social intelligence. *arXiv preprint arXiv:2508.03905*, 2025.

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? In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 57654–57689. Curran Associates, Inc., 2025. doi: 10.52202/085713-1933. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/537d5aa768c2d534016a4d06f87bc8fb-Paper-Conference.pdf.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, volume 2024, pp. 40975–41019, 2024.

APPENDIX GUIDE

Use the guide below to find the experimental procedures, full results, and supporting analyses. Several tables break down the same observations by game, model, or payoff setting; they are different views of the data, not additional experiments.

Evidence group	Reading path
Design and measurement	Statistical units and directions (Appendix A); checkpoints, games, SFT sources, and grids (B–E); executed prompts and retries (M–N).
Complete canonical results	Fitted responses, history contrasts, configuration and lineage views (F–H, O); Trust and older-record probes (I–J); every canonical configuration (Q).
Fixed-input evidence	PGG history visibility: rates, interactions and intervals (K); seven-game future instructions (L); fixed-archive estimation and calibration (P).
Interpretation and case studies	Realized payoff identities and repeated-game benchmark (R); related work (S); seven-stage Nemotron case (T); programmed-partner action values (U).

A Reproducibility and Statistical Units

The primary dataset contains 52,920 unique replicate aggregates: 21 checkpoints × 84 configurations × (20 one-shot draws + two history lengths × five trajectory seeds), yielding 5,292 checkpoint-by-configuration-by-condition cells. The original 82-configuration panel contains 51,660 replicate aggregates and 5,166 checkpoint-by-configuration-by-condition cells. Appendix E explains the payoff-grid changes and their 90-configuration union. Its full-window aggregate data are retained for the original-grid sensitivity analysis; new CPR cells additionally retain raw actions. The reported main panel contains a completed aggregate for every declared replicate key. Failed requests and interrupted prefixes are retained as execution records and are not counted as additional replicates or assigned zero cooperation. Appendix N reports the separate retry and completion-bound analysis.

Stage differences, history changes, and fitted payoff responses are defined in Appendices A.1 and F. Within a checkpoint and configuration, SDs compare five repeated runs or 20 one-shot draws. Maintable SDs instead compare the seven lineage estimates after averaging each lineage's runs and configurations. Tables 2 and 16 report means; Tables 17–22 show each lineage separately. For a difference, we calculate its SD from paired lineage estimates. We keep each repeated trajectory intact rather than treating its rounds as independent runs.

We chose these seven lineages to obtain released training-stage comparisons, rather than sampling them randomly from all language models. The exploratory t_6 intervals and leave-one-lineage-out results show how summaries depend on differences among these models. The PGG history-visibility experiment shows or hides saved interaction records (Section 5.2.2; Appendix K). Its sampling intervals describe how estimates would vary with newly sampled records and answers from the same checkpoints and recorded histories.

A.1 Measurement notation and comparisons

Observations and rates. A replicate b contains all n players' decisions over $T = 200$ repeated rounds or $T = 1$ one-shot round. Let $Y_{bti} \in \{0, 1\}$ indicate player i's cooperative action in round t; for TD it is the normalized claim in $[0, 1]$. The replicate rate is $\hat{r}_b = (Tn)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{n} Y_{bti}$. Averaging replicates gives $\hat{R}^e_{\ell gcs}$ with expectation $R^e_{\ell gcs}$. Here ℓ indexes lineage, g game, c complete payoff configuration, $s \in \{B, S, A\}$ training stage, and $e \in \{\mathrm{OS}, \mathrm{HL2}, \mathrm{HL40}\}$ one-shot or repeated-history condition. Rates are fractions and are displayed as percentages.

Training-stage and history comparisons. With fixed lineage, game, configuration, and evaluation condition suppressed, the stage differences are $D_{S/B} = R_S - R_B$, $D_{A/B} = R_A - R_B$, and $D_{A/S} = R_A - R_S$. For repeated play, $R_s(L)$ denotes the expected rate at stage s with at most L completed rounds visible. The main text writes this rate as $p_s(L) = R_s(L)$ to distinguish it from Donation-PD's payoff ratio R. As in Equation 1, $H_s = R_s(40) - R_s(2)$ and $K_{s/t} = H_s - H_t$. Thus $K_{S/B}$, $K_{A/B}$, and $K_{A/S}$ all compare the same kind of history-related change between two stages, and $K_{A/B} = K_{S/B} + K_{A/S}$. Reports multiply D, H, and K by 100 for pp. We form each contrast within lineage before calculating its cross-lineage SD.

Payoff-response coordinates. Holding checkpoint, evaluation condition, and other payoff parameters fixed, let x_j be the raw payoff parameter and $y_j \in [0, 1]$ its measured rate at tested point $j = 1, \ldots, J_g$, where J_g is the number of tested values. The axis direction $q_g \in \{-1, +1\}$ sets the normalized coordinate:

$$z_j = \begin{cases} (x_j - x_{\min})/(x_{\max} - x_{\min}), & q_g = +1, \\ (x_{\max} - x_j)/(x_{\max} - x_{\min}), & q_g = -1. \end{cases}$$

We use $q_g = -1$ for PGG's MPCR, Donation-PD's return ratio, and Assurance's success premium; $q_g = +1$ for CPR's extraction parameter, Volunteer's cost ratio, and TD's reward–penalty parameter. Hence the reported axis runs toward higher cooperation costs, lower cooperation benefits, or a larger reward for noncooperation, as appropriate for the game. Fit $\hat{f}(z) = \hat{\eta} + \hat{\beta}_z z$ by ordinary least squares (OLS) to all tested points, where $\hat{\eta}$ is the fitted intercept and $\hat{\beta}_z$ the fitted slope. The fitted drop in percentage points is

$$S = 100[\hat{f}(0) - \hat{f}(1)] = -100\hat{\beta}_z. \tag{2}$$

Subscripts B, S, and A distinguish the Base, SFT, and Aligned fitted drops. All points and their numerical spacing enter the fit. A positive fitted drop describes a decrease across the tested span, without requiring every neighboring pair of observations to decrease. Appendix F gives raw-axis equivalents and payoff-response changes with history.

One-shot versus repeated play. Let $\mathcal{C}_g$ be the game's configuration set and w_{gc} its fixed configuration weights, summing to one. Within each lineage and game, $\bar{R}^e_{\ell gs} = \sum_{c\in\mathcal{C}_g} w_{gc}\hat{R}^e_{\ell gcs}$. Section 5.1 compares the Aligned-minus-SFT differences

$$\delta^e_{\ell g} = 100(\bar{R}^e_{\ell gA} - \bar{R}^e_{\ell gS}), \qquad G_{\ell g} = \delta^{\mathrm{HL40}}_{\ell g} - \delta^{\mathrm{OS}}_{\ell g}. \tag{3}$$

Both are already in pp. Negative $G_{\ell g}$ means that the Aligned-minus-SFT difference is lower in repeated play than in one-shot decisions. Table 4 reports the means and SDs of these lineage-level estimates for each game.

Game groups. Cross-game summaries weight games equally. Cooperation-level and history summaries use the six binary games, including Trust. The five symmetric binary games are PGG, CPR, Donation-PD, Volunteer, and Assurance. Payoff-response summaries exclude Trust's distinct role-specific incentives; six-symmetric-game summaries additionally include TD's normalized-claim response. TD is not included in the binary cooperation averages.

A.2 Aggregation and the meaning of dispersion

For each lineage and game, we average runs within each complete payoff configuration, then combine configurations with the stated weights. We average players within each run before combining configurations. A five-player configuration therefore does not receive more weight simply because it has more players. Let $x_{\ell g}$ be the resulting estimate for lineage ℓ in game g, such as a cooperation rate or a difference between stages or history lengths. The mean and SD across the seven lineages are

$$\bar{x}_g = \frac{1}{7}\sum_{\ell=1}^{7} x_{\ell g}, \qquad s_{\mathrm{lin},g} = \sqrt{\frac{1}{6}\sum_{\ell=1}^{7}(x_{\ell g} - \bar{x}_g)^2}. \tag{4}$$

The exploratory interval $\bar{x}_g \pm t_{0.975,6}\, s_{\mathrm{lin},g}/\sqrt{7}$ uses differences among these seven estimates, where $t_{0.975,6}$ is the 97.5th percentile of Student's t distribution with six degrees of freedom. We call this a model-composition interval because it depends on which lineages are included. The archived-record experiment also estimates uncertainty from sampling new records and answers while keeping the models fixed (Appendix P). For the one-shot-to-repeated comparison, $x_{\ell g} = G_{\ell g} = \delta^{\mathrm{HL40}}_{\ell g} - \delta^{\mathrm{OS}}_{\ell g}$. Its SD is calculated from these seven differences and includes their within-lineage covariance:

$$s^2_G = s^2_{\delta^{\mathrm{HL40}}} + s^2_{\delta^{\mathrm{OS}}} - 2\widehat{\mathrm{Cov}}_\ell(\delta^{\mathrm{HL40}}, \delta^{\mathrm{OS}}). \tag{5}$$

Here s_{δ^e} is the SD of the seven stage differences under condition e, and $\widehat{\mathrm{Cov}}_\ell$ is their covariance across lineages. The one-shot and repeated rates can each vary widely across models while their difference varies less: a lineage that is high in one condition may also be high in the other. We therefore calculate the SD of paired differences directly, rather than subtracting or adding the two SDs. Large SDs can reflect changes of different sizes or changes in opposite directions; the direction counts distinguish these cases.

The captions identify what varies in each SD. Table 4 compares seven lineages after averaging each lineage's configurations and runs. Tables 17–22 instead compare five runs of the same checkpoint. In Table 28, each lineage row compares four payoff settings, while the final row compares seven lineages. These SDs include sampling noise in the estimated rates. Bounds for invalid decisions answer another question: how much could the result change if every missing decision took either legal action?

For one-shot versus repeated play, we pair the two condition means within each lineage, not individual one-shot answers with individual repeated runs. We then compute the SD of the seven differences.

A.3 Effect size, dispersion, and directional coverage

Direction counts depend on what is being counted. The 7/7 and 6/7 counts first average games and payoffs within each lineage, then count lineages. The 66/68 and 64/68 counts instead average the seven lineages at each binary-game configuration, then count configurations. The finer cell counts keep every lineage and configuration separate. The table below gives the denominator and source for each count.

Finding	Direction and unit	Coverage	Existing source
Cooperation	SFT–Base below zero, binary configuration means	66/68	Appendix D
	Aligned–Base below zero, binary configuration means	64/68	Appendix D
Payoff response	Stage–Base fitted response above zero, symmetric binary game means	5/5 each	Table 2
History change	$H_S < 0$ and $H_A < 0$, six-game lineage summaries	7/7 each	Table 12
Stage contrast	$K_{S/B} < 0$ and $K_{A/B} < 0$, six-game lineage summaries	6/7 each	Table 12
Cell-level detail	$K_{S/B} < 0$; $K_{A/B} < 0$, binary lineage-by-config cells	321/476; 336/476	Sum of binary rows in Table 14
Interaction setting	$G < 0$, symmetric binary game means	5/5	Table 4
	$G < 0$, lineages within PGG/CPR/Donation-PD/Volunteer/Assurance	5/7; 5/7; 4/7; 7/7; 3/7	Table 4
Shared records	$I < 0$, lineage estimates	5/7	Table 28
	$H^{\text{vis}}_{\text{CONTINUE}} < 0$; $H^{\text{vis}}_{\text{FINAL}} > 0$, lineage estimates	5/7; 3/7	Table 28

Different sources of spread. In Table 28, the CONTINUE history contrast is -14.87 pp for DeepSeek-Math but $+9.00$ pp for Nemotron-Cascade. Within each model, the four-payoff SD is much smaller: 2.72 and 1.32 pp, respectively. A model can therefore respond similarly across its payoff settings while responding very differently from another model. The SD across models captures that difference together with sampling noise.

Interpreting dense plots. The 476 binary cells in Table 14 are seven lineages evaluated at 68 payoff settings, not 476 independent models. Counting cells gives more weight to games with more configurations. The main history average instead gives each game equal weight, so the two summaries need not agree exactly.

One-shot-to-repeated lineage directions. For $G = \delta^{\text{HL40}} - \delta^{\text{OS}}$, the five symmetric binary-game counts appear above. TD has 5/7 negative lineage estimates and Trust 2/7. Table 4 reports every game's paired mean, SD, and count.

B Model Checkpoints and Training Ancestry

Table 7 maps the display names in Section 2 to the 21 checkpoints used by the primary panel: one Base, one SFT, and one Aligned checkpoint per lineage. Repository identifiers retain their official spelling; a colon separates a Hugging Face repository from its checkpoint subfolder. In particular, OLMo-3-7B Base denotes `allenai/Olmo-3-1025-7B`; `1025` is part of the release identifier, not a training-stage label. The Aligned objective is specified for each lineage. All three stages within a triplet have the parameter scale shown in the lineage column.

Table 7: All 21 checkpoints in the primary panel: one Base, SFT, and Aligned release per lineage. Identifiers link to the Hugging Face repositories; Nemotron's SFT and RLHF suffixes identify checkpoint subdirectories. Aligned training objectives appear in parentheses.

Lineage	Stage	Hugging Face repository
OLMo-3-7B	Base	`allenai/Olmo-3-1025-7B`
OLMo-3-7B	SFT	`allenai/Olmo-3-7B-Instruct-SFT`
OLMo-3-7B	Aligned (DPO)	`allenai/Olmo-3-7B-Instruct-DPO`
SmolLM2-1.7B	Base	`HuggingFaceTB/SmolLM2-1.7B`
SmolLM2-1.7B	SFT	`HuggingFaceTB/SmolLM2-1.7B-sft-only`
SmolLM2-1.7B	Aligned (DPO)	`HuggingFaceTB/SmolLM2-1.7B-Instruct`
Mistral/Zephyr-7B	Base	`mistralai/Mistral-7B-v0.1`
Mistral/Zephyr-7B	SFT	`HuggingFaceH4/mistral-7b-sft-beta`
Mistral/Zephyr-7B	Aligned (DPO)	`HuggingFaceH4/zephyr-7b-beta`
Llama/Tülu-3-8B	Base	`meta-llama/Llama-3.1-8B`
Llama/Tülu-3-8B	SFT	`allenai/Llama-3.1-Tulu-3-8B-SFT`
Llama/Tülu-3-8B	Aligned (DPO)	`allenai/Llama-3.1-Tulu-3-8B-DPO`
InternLM2-7B	Base	`internlm/internlm2-base-7b`
InternLM2-7B	SFT	`internlm/internlm2-chat-7b-sft`
InternLM2-7B	Aligned (online-RLHF)	`internlm/internlm2-chat-7b`
DeepSeekMath-7B	Base	`deepseek-ai/deepseek-math-7b-base`
DeepSeekMath-7B	SFT	`deepseek-ai/deepseek-math-7b-instruct`
DeepSeekMath-7B	Aligned (GRPO/RL)	`deepseek-ai/deepseek-math-7b-rl`
Nemotron-Cascade-8B	Base	`Qwen/Qwen3-8B-Base`
Nemotron-Cascade-8B	SFT	`nvidia/Nemotron-Cascade-8B-Intermediate-ckpts:SFT`
Nemotron-Cascade-8B	Aligned (RLHF)	`nvidia/Nemotron-Cascade-8B-Intermediate-ckpts:RLHF`

Which post-SFT release is Aligned? OLMo-3, SmolLM2, Zephyr, and Tülu-3 use the DPO releases in the table. InternLM2 uses its online-RLHF Chat release, DeepSeekMath its GRPO-trained RL release, and Nemotron-Cascade the `RLHF` subfolder of the intermediate-checkpoint repository. In particular, OLMo-3 and Tülu-3 use DPO rather than their publishers' later reinforcement learning with verifiable rewards (RLVR) releases. The primary Nemotron triplet ends at RLHF; Appendix T adds instruction-following RL (IFRL), mathematical RL (MathRL), code RL (CodeRL), and software-engineering RL (SWE-RL) to form the seven-stage case study.

Training recipes across lineages. The stage labels identify training procedures, while the demonstrations and optimization objectives differ across lineages. SFT already shapes desired behavior through its target responses. For example, the Mistral-SFT-Beta model card describes supervised training on filtered UltraChat dialogues generated by ChatGPT before the DPO step used for Zephyr. The Tülu-3-SFT model card describes a mixture of public synthetic and human-created datasets, followed by a separate DPO release. Together with the different post-SFT objectives in Table 7, these recipes motivate comparing each released training transition rather than interpreting SFT and Aligned as uniform amounts of behavioral alignment. Recipe differences are one possible contributor to the heterogeneous Aligned-minus-SFT results in Section 5.1.

Base and ancestry details. The InternLM2 SFT model card documents training from InternLM2-Base; its chat checkpoint follows SFT with online RLHF. We therefore use `internlm/internlm2-base-7b`. The separately named `internlm/internlm2-7b` includes additional capability/domain pretraining and is not an interchangeable Base alias. The SmolLM2 SFT card names a SmolLM2-1.7B-8k parent, whereas the evaluated Base is the published 1.7B endpoint listed in the table; Appendix M reports the corresponding no-SmolLM2 sensitivity. Appendix M describes deployment and available raw-log coverage.

C Games, Metrics, and Complete Parameter Grids

Here u_i is player i's one-round payoff in game points. In binary-action games, C denotes the designated cooperative action ($a_i = 1$) and D the alternative ($a_i = 0$); these are explanatory labels for the implemented actions. We write $\mathbf{1}\{E\}$ for the indicator of event E: it equals one when the event holds and zero otherwise. Parameters are local to each game. Every numerical example below uses a configuration in the amended main grid.

Reading the matrices. For the first five games, rows give a focal player's action and columns give $m_{-i} = \sum_{j \neq i} a_j$, the number of other players choosing C. Each entry is that focal player's payoff, not a group total. These symmetric games depend on the other players only through this count, so the compact matrices specify every player's payoff for every joint action without treating a multiplayer game as a two-player game. TD entries instead give the ordered pair (u_1, u_2), and Trust entries give (u_1, u_2, u_3). The entries are deterministic payoffs, not empirical means or cooperation rates.

C.1 Public-goods game (PGG)

Each of n players contributes its unit endowment (C) or retains it (D). For $k = \sum_j a_j$ contributors and MPCR $\alpha = r/n$, the payoff is $u_i = 1 - a_i + \alpha k$. We cross $n \in \{3, 5\}$ with $\alpha \in \{0.4, 0.5, 0.6, 0.8\}$ in the amended main panel. A contribution increases the return to every player by α, while its contributor gives up the unit endowment; its net private cost is $1 - \alpha$.

General payoff matrix: focal payoff u_i.

Own action	$m_{-i} = 0$	$m_{-i} = 1$	$\cdots$	$m_{-i} = n - 1$
C: contribute	α	2α	$\cdots$	$n\alpha$
D: retain	1	$1 + \alpha$	$\cdots$	$1 + (n-1)\alpha$

Example: $n = 3$, $\alpha = 0.6$, hence $r = 1.8$.

Own action	$m_{-i} = 0$	$m_{-i} = 1$	$m_{-i} = 2$
C: contribute	0.6	1.2	1.8
D: retain	1.0	1.6	2.2

For example, one contributor and two noncontributors receive $(0.6, 1.6, 1.6)$ when the contributor is Player 1. Original MPCR 0.7 observations remain in the archived 82-grid sensitivity; the four-point main fit excludes them at both player counts and every checkpoint/exposure consistently.

C.2 Common-pool resource game (CPR)

Three players choose restrained extraction $x_i = 1$ (C) or high extraction $x_i = h$ (D), with $h \in \{1.25, 1.5, 2.5, 3\}$. Their payoff is $u_i = x_i(V - \sum_j x_j)$, where $V = 3h + 1 + 2\theta$ and the amended main grid is $\theta \in \{0.2, 0.4, 0.6, 0.8\}$. Thus m_{-i} counts the other restrained extractors. The resource-return rule is applied anew each round; there is no physical resource stock carried across rounds.

General payoff matrix: focal payoff u_i.

Own action	$m_{-i} = 0$	$m_{-i} = 1$	$m_{-i} = 2$
C: extract 1	$V - 1 - 2h$	$V - 2 - h$	$V - 3$
D: extract h	$h(V - 3h)$	$h(V - 1 - 2h)$	$h(V - 2 - h)$

Example: $h = 1.5$, $\theta = 0.4$, hence $V = 6.3$.

Own action	$m_{-i} = 0$	$m_{-i} = 1$	$m_{-i} = 2$
C: extract 1	2.30	2.80	3.30
D: extract 1.5	2.70	3.45	4.20

In the example, all restrained extractors earn 3.30 each, while all high extractors earn 2.70 each. Holding the other two actions fixed, high extraction nevertheless gives the larger private payoff in every column.

The separately collected CPR 0.4/0.6 points and original 0.5 archive are described in Appendices E and N.

C.3 Donation Prisoner's Dilemma (Donation-PD)

Each of three players starts with 2 points. Cooperating pays cost c to give benefit b to each other player; defecting pays no cost and gives nothing. The payoff is $u_i = 2 - ca_i + b\sum_{j \neq i} a_j$. We cross $c \in \{0.2, 0.6, 1.2\}$ with $R = 2b/c \in \{1.1, 1.4, 2.0, 2.4\}$. At fixed c, changing R changes others' contribution returns while the immediate private difference remains $u_D - u_C = c$. Its oriented response thus measures sensitivity to external returns.

General payoff matrix: focal payoff u_i.

Own action	$m_{-i} = 0$	$m_{-i} = 1$	$m_{-i} = 2$
C: donate	$2 - c$	$2 - c + b$	$2 - c + 2b$
D: do not donate	2	$2 + b$	$2 + 2b$

Example: $c = 0.6$, $R = 2$, hence $b = 0.6$.

Own action	$m_{-i} = 0$	$m_{-i} = 1$	$m_{-i} = 2$
C: donate	1.4	2.0	2.6
D: do not donate	2.0	2.6	3.2

C.4 Volunteer's Dilemma (Volunteer)

Three players choose whether to volunteer. If at least one volunteers, *every* player receives benefit B, and *each* volunteer separately pays cost c. If nobody volunteers, all receive zero. Thus $u_i = B\mathbf{1}\{\sum_j a_j \geq 1\} - ca_i$, with $B \in \{2, 4, 8\}$ and $\rho = c/B \in \{0.125, 0.375, 0.625, 0.875\}$.

A volunteer always earns the net payoff $B - c = B(1 - \rho)$, whether or not anyone else volunteers. A nonvolunteer earns B if another player volunteers and zero otherwise. The benefit is paid to each player, not divided among players; the cost is paid by every volunteer, not shared among volunteers. Across the four tested cost ratios, a volunteer therefore receives $0.875B$, $0.625B$, $0.375B$, or $0.125B$, respectively.

General payoff matrix: focal payoff u_i.

Own action	$m_{-i} = 0$	$m_{-i} = 1$	$m_{-i} = 2$
C: volunteer	$B - c$	$B - c$	$B - c$
D: do not volunteer	0	B	B

Example: $B = 4$, $\rho = 0.375$, hence $c = 1.5$.

Own action	$m_{-i} = 0$	$m_{-i} = 1$	$m_{-i} = 2$
C: volunteer	2.5	2.5	2.5
D: do not volunteer	0	4	4

For example, (C, D, D) gives $(2.5, 4, 4)$, (C, C, D) gives $(2.5, 2.5, 4)$, and (D, D, D) gives $(0, 0, 0)$. Volunteering is privately beneficial when nobody else volunteers; otherwise, not volunteering avoids the cost.

Volunteer is a threshold-one public-good dilemma: everyone benefits from provision, but each prefers another player to bear its cost. Unlike a dominant-defection dilemma, the best response depends on whether another player volunteers. Since $0 < c < B$ throughout our grid, the pure-strategy Nash equilibria have exactly one volunteer. The cooperation rate here measures willingness to supply the benefit; group performance depends on securing provision while avoiding redundant costs (Appendix R.1).

C.5 Assurance

Three players choose to join a project (C) or take a safe outside option (D). A contributor receives h if at least $k \in \{2, 3\}$ players, including itself, contribute, and zero otherwise. Opting out always gives 3, including when the others complete the project. Thus

$$u_i = \begin{cases} h\mathbf{1}\{m_{-i} + 1 \geq k\}, & a_i = 1, \\ 3, & a_i = 0, \end{cases} \qquad h = 3(1 + \lambda), \quad \lambda \in \{0.05, 0.10, 0.20, 0.40\}.$$

Here h is the contributor's full successful-project payoff, not a bonus added to 3 and not a shared pool.

General payoff matrix: focal payoff u_i, $k \in \{2, 3\}$.

Own action	$m_{-i} = 0$	$m_{-i} = 1$	$m_{-i} = 2$
C: join	0	$h\mathbf{1}\{k = 2\}$	h
D: safe option	3	3	3

Example: $k = 2$, $\lambda = 0.2$, hence $h = 3.6$.

Own action	$m_{-i} = 0$	$m_{-i} = 1$	$m_{-i} = 2$
C: join	0	3.6	3.6
D: safe option	3	3	3

In this example, (C, C, D) gives $(3.6, 3.6, 3)$, whereas (C, D, D) gives $(0, 3, 3)$. At threshold $k = 3$ with the same λ, the contributor's middle-column payoff becomes zero; the other entries are unchanged.

Full cooperation is a strict stage-game Nash equilibrium throughout this grid: against two contributors, joining pays $h > 3$, more than the outside option. All opting out is also a strict equilibrium because one participant cannot meet either threshold. The cooperative outcome makes every player better off, but reaching it requires coordinating participation rather than overcoming a profitable deviation from full cooperation (Appendix R).

C.6 Traveler's Dilemma (TD)

Two players choose integer claims from 2 to $K + 1$, with $K \in \{8, 16, 32, 64\}$ legal levels. Equal claims pay that claim to both players. For unequal claims, the lower claimant receives its claim plus bonus B, while the higher claimant receives the lower claim minus B, where $B = 1 + \tau(K - 1)$ and $\tau \in \{0.025, 0.05, 0.10, 0.20\}$. For a chosen claim c, the reported behavioral outcome is $(c - 2)/(K - 1)$; this normalized claim is distinct from payoff.

General payoff matrix: entries (u_1, u_2).
Rows are Player 1's claims; columns are Player 2's claims.

Claim	2	3	$\cdots$	$K+1$
2	$(2, 2)$	$(2+B, 2-B)$	$\cdots$	$(2+B, 2-B)$
3	$(2-B, 2+B)$	$(3, 3)$	$\cdots$	$(3+B, 3-B)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$K+1$	$(2-B, 2+B)$	$(3-B, 3+B)$	$\cdots$	$(K+1, K+1)$

For $K = 8$ and $\tau = 0.1$, $B = 1.7$: equal claims of 9 pay $(9, 9)$, while claims $(8, 9)$ pay $(9.7, 6.3)$. Undercutting raises the lower claimant's payoff by 0.7, and the general matrix specifies all 64 joint actions.

C.7 TRUST

Player 1 is an investor with endowment $E = 10$; Players 2 and 3 are trustees. Keeping the endowment gives $(10, 0, 0)$. Investing creates $G = 5m$ for each trustee, with multiplier $m \in \{2, 3, 4\}$. Each trustee's contingent action either returns $r = 5(1 + s)$ to the investor (C) or retains everything (D), with $s \in \{0.2, 0.4, 0.6, 0.8\}$. Returning therefore leaves a trustee $G - r$, while retaining gives G, conditional on investment. The investor receives the sum returned by the two trustees.

General payoff matrix: entries (u_1, u_2, u_3).
Columns give the ordered trustee actions (a_2, a_3).

Investor action	(C, C)	(C, D)	(D, C)	(D, D)
C: invest	$(2r, G-r, G-r)$	$(r, G-r, G)$	$(r, G, G-r)$	$(0, G, G)$
D: keep	$(10, 0, 0)$	$(10, 0, 0)$	$(10, 0, 0)$	$(10, 0, 0)$

Example: $m = 3$, $s = 0.4$, hence $G = 15$ and $r = 7$.

Investor action	(C, C)	(C, D)	(D, C)	(D, D)
C: invest	$(14, 8, 8)$	$(7, 8, 15)$	$(7, 15, 8)$	$(0, 15, 15)$
D: keep	$(10, 0, 0)$	$(10, 0, 0)$	$(10, 0, 0)$	$(10, 0, 0)$

The implementation elicits all three intentions simultaneously; trustee returns are executed only if the investor invests. Thus a trustee's recorded cooperative intention need not produce an actual transfer when the investor keeps the endowment. Appendix I separates these role-specific measurements.

D STAGE COMPARISONS AND THEIR SFT REFERENCE

At HL40, the configuration-mean SFT–Base and Aligned–Base contrasts are negative in 66 and 64 of the 68 binary configurations, respectively. Both contrasts have negative means in each of the six binary games. For OLMo and SmolLM2, the earlier Base–SFT and SFT–Aligned comparisons used separately collected SFT runs. The other five lineages used the same repeated SFT runs in both comparisons. Our main three-stage comparisons use one SFT dataset per lineage: the one used for the original table of absolute cooperation rates.

We match payoff configurations and run IDs, average the full recorded window, and then calculate differences. For OLMo and SmolLM2, this means recomputing Aligned minus SFT using the same SFT dataset as the Base comparison. Matching IDs specifies which runs are subtracted; separately collected runs need not share the same random draws.

Table 8: Stage changes by game in the main 84-configuration panel (pp): mean ± SD over seven lineages, after averaging configurations and replicate IDs within lineage. A denotes the Aligned checkpoint defined in Section 2. Both stage contrasts use the same SFT reference, described in Appendix D. TD uses normalized claims; Trust averages all-seat intentions, separated by role in Appendix I.

Game	Contrast	One-shot	HL2	HL40
PGG	SFT–Base	-27.50 ± 18.85	-17.23 ± 34.41	-20.44 ± 40.99
	A–SFT	$+8.62 \pm 6.78$	$+3.88 \pm 15.39$	-0.77 ± 20.44
CPR	SFT–Base	-32.29 ± 20.36	-22.38 ± 29.62	-34.34 ± 29.49
	A–SFT	$+3.47 \pm 22.35$	-2.31 ± 15.20	-7.08 ± 14.96
Donation-PD	SFT–Base	-38.08 ± 39.11	-33.49 ± 37.38	-39.97 ± 41.73
	A–SFT	-1.71 ± 7.11	-1.18 ± 5.47	-4.21 ± 5.51
Volunteer	SFT–Base	-17.42 ± 26.54	-14.80 ± 21.76	-28.64 ± 29.60
	A–SFT	$+2.66 \pm 14.47$	-6.82 ± 13.79	-7.25 ± 11.63
Assurance	SFT–Base	-15.77 ± 28.26	-11.53 ± 15.80	-10.35 ± 21.13
	A–SFT	$+1.93 \pm 13.67$	$+7.66 \pm 16.81$	$+1.59 \pm 8.14$
TD	SFT–Base	$+21.44 \pm 17.35$	$+28.32 \pm 37.50$	$+27.65 \pm 31.33$
	A–SFT	$+2.48 \pm 9.71$	-3.94 ± 7.17	-5.50 ± 8.82
Trust	SFT–Base	-16.11 ± 17.60	-19.08 ± 14.60	-26.38 ± 23.36
	A–SFT	-3.91 ± 8.47	-1.82 ± 9.77	$+0.21 \pm 12.99$

E Original-Grid Sensitivity and the 84-Config Amendment

Sections 3, 4, and 5.1 use the same 84-configuration grid for one-shot, HL2, and HL40 results. The original 82 configurations comprised 10 PGG, 12 CPR, 12 Donation-PD, 12 Volunteer, 8 Assurance, 16 TD, and 12 Trust settings. To show four values per main parameter, we removed PGG MPCR 0.7 at both player counts and replaced CPR $\theta = 0.5$ at four h values with $\theta = 0.4, 0.6$ at those same values. Thus $82 - 2 - 4 + 8 = 84$, with 8 PGG and 16 CPR configurations. This change was made after the original experiment. A configuration fixes all payoff parameters; the count does not multiply by the number of models or stages. All other complete configurations remain identical. The original 0.5 CPR and 0.7 PGG points enter the grid-sensitivity analyses. The table below compares the grids using the same SFT reference and weights, with changes in percentage points. The new CPR points were collected in a later deployment, so this comparison includes both the grid change and that deployment difference. Separate tests have their own payoff settings: the continuation test uses the original 82 payoffs (94 settings after separating focal roles); Appendix T uses the same 82-grid for all seven Nemotron checkpoints; the history-visibility experiment in Appendix K uses four PGG settings; and Appendix U uses 16 configurations.

The original 82 and amended 84 share 76 configurations (47,880 replicate rows), whose values are exactly unchanged. The eight new CPR configurations contribute 5,040 rows; six old configurations not in the main grid contribute 3,780. Their complete union has 90 configurations and 56,700 unique rows. The 80 resumed trajectories retain their original replicate identities. These three grids cover the original design, amended main panel, and complete observed union. Table 2 reports the main four-point means and fitted drops; Table 16 adds mean paired stage differences. Table 9 compares all three grid summaries.

Table 9: Original 82, amended main 84, and complete observed union 90 grids. Each cell gives mean ± seven-lineage SD (pp), followed on the next line by the existing t_6 model-composition interval. H/K use six binary games; $S/T/Q$ use six symmetric games, including TD. The union is a reweighting of existing observations, not an independent replication.

Metric	Original 82	Main 84	Observed 90
H_B	-1.26 ± 3.97	-1.17 ± 3.92	-1.20 ± 3.89
	$[-4.93, +2.41]$	$[-4.79, +2.45]$	$[-4.79, +2.40]$
H_S	-8.22 ± 5.67	-8.11 ± 5.46	-8.05 ± 5.52
	$[-13.47, -2.97]$	$[-13.16, -3.05]$	$[-13.16, -2.95]$
H_A	-10.86 ± 6.46	-10.93 ± 6.40	-10.88 ± 6.49
	$[-16.83, -4.88]$	$[-16.85, -5.01]$	$[-16.88, -4.88]$
$K_{S/B}$	-6.96 ± 7.57	-6.93 ± 7.30	-6.86 ± 7.37
	$[-13.96, +0.04]$	$[-13.68, -0.19]$	$[-13.67, -0.04]$
$K_{A/B}$	-9.59 ± 7.38	-9.76 ± 7.47	-9.69 ± 7.49
	$[-16.42, -2.77]$	$[-16.67, -2.85]$	$[-16.61, -2.76]$
$K_{A/S}$	-2.64 ± 4.58	-2.82 ± 4.37	-2.83 ± 4.58
	$[-6.87, +1.60]$	$[-6.87, +1.22]$	$[-7.06, +1.41]$
$S_S - S_B$, OS	$+1.46 \pm 1.42$	$+1.49 \pm 1.42$	$+1.48 \pm 1.34$
	$[+0.15, +2.78]$	$[+0.18, +2.80]$	$[+0.24, +2.72]$
$S_S - S_B$, HL2	$+1.72 \pm 1.77$	$+1.76 \pm 1.72$	$+1.71 \pm 1.84$
	$[+0.09, +3.35]$	$[+0.17, +3.35]$	$[+0.00, +3.41]$
$S_S - S_B$, HL40	$+2.97 \pm 4.91$	$+2.83 \pm 5.05$	$+3.00 \pm 4.83$
	$[-1.56, +7.51]$	$[-1.84, +7.50]$	$[-1.46, +7.47]$
T_S	$+2.61 \pm 2.63$	$+2.47 \pm 2.51$	$+2.66 \pm 2.65$
	$[+0.18, +5.03]$	$[+0.14, +4.79]$	$[+0.21, +5.11]$
T_A	$+3.68 \pm 2.67$	$+3.74 \pm 2.86$	$+3.68 \pm 2.65$
	$[+1.21, +6.15]$	$[+1.10, +6.38]$	$[+1.22, +6.13]$
$Q_{S/B}$	$+1.25 \pm 4.10$	$+1.07 \pm 4.09$	$+1.29 \pm 3.94$
	$[-2.54, +5.04]$	$[-2.71, +4.85]$	$[-2.35, +4.94]$
$Q_{A/B}$	$+2.33 \pm 4.58$	$+2.34 \pm 4.76$	$+2.31 \pm 4.47$
	$[-1.91, +6.56]$	$[-2.07, +6.75]$	$[-1.82, +6.45]$

F PAYOFF FITS AND HISTORY COMPARISONS

Across the six symmetric games, SFT–Base responsiveness is 1.49 ± 1.42 one-shot, 1.76 ± 1.72 at HL2, and 2.83 ± 5.05 at HL40. The long-history contrast is largest; the tables below retain lineage directions, dispersion, and game/stage profiles.

Raw-parameter and normalized-coordinate slopes. For one fixed lineage, training stage, interaction condition, and combination of the other game parameters, let x_j be a tested raw parameter value and y_j the associated cooperation-rate estimate on the $[0, 1]$ scale. A complete payoff configuration consists of x_j together with those fixed other parameters. The fit gives equal weight to all J_g observed primary-axis values. The amended main grid has four values per axis; the historical grid retains its actual three- or five-point axes where applicable. Let $\bar{x} = J_g^{-1} \sum_j x_j$ and $\bar{y} = J_g^{-1} \sum_j y_j$. With an intercept, the OLS slope on the raw parameter is

$$\hat{\beta}_x = \frac{\sum_j (x_j - \bar{x})(y_j - \bar{y})}{\sum_j (x_j - \bar{x})^2}, \qquad \hat{\eta}_x = \bar{y} - \hat{\beta}_x \bar{x}.$$

The normalized coordinate in Appendix A.1 has x increasing with z when $q_g = +1$ and decreasing with z when $q_g = -1$. Therefore $\hat{\beta}_z = q_g(x_{\max} - x_{\min})\hat{\beta}_x$, and Equation 2 is equivalently

$$S = -100 q_g \hat{\beta}_x (x_{\max} - x_{\min}).$$

The factor 100 converts a rate difference into pp. The range normalization expresses a fitted response across the tested range of one game, whose economic meaning is specified in Table 1. Different games vary different private costs, external returns, or coordination rewards. For example, increasing Donation-PD's return ratio at fixed contribution cost changes the benefit to the other players; it leaves the current private contribution cost unchanged.

History changes in the fitted response. With $S_s(L)$ denoting the fitted response for stage s and history limit L, define $T_s = S_s(40) - S_s(2)$, $Q_{S/B} = T_S - T_B$, and $Q_{A/B} = T_A - T_B$. A positive $Q_{S/B}$ means that increasing history raises the fitted response more for SFT than for Base. These quantities inherit the pp scale of S. Fit each replicate on its complete parameter grid and then apply the existing within-lineage, secondary-parameter, and game weights; the shared grid and OLS linearity give the same point estimate as fitting the corresponding averaged rates. For the six-game

Table 10: All-point oriented payoff responsiveness S (pp over the observed parameter span); six symmetric games equally weighted within lineage, then mean $\pm$ SD over seven lineages. Displayed stage differences equal differences of stage means before rounding.

Exposure	Base (S_B)	SFT (S_S)	Aligned (S_A)	$S_S - S_B$	$S_A - S_S$
one-shot	0.46 ± 1.09	1.94 ± 1.26	3.69 ± 2.07	$+1.49 \pm 1.42$	$+1.75 \pm 1.99$
HL2	1.40 ± 2.55	3.16 ± 1.53	3.58 ± 1.97	$+1.76 \pm 1.72$	$+0.42 \pm 1.44$
HL40	2.79 ± 5.90	5.62 ± 3.09	7.32 ± 4.27	$+2.83 \pm 5.05$	$+1.70 \pm 2.17$

average including TD, the HL40 SFT–Base response interval includes zero. The intervals for Q are also broad: models differ in how much longer history changes their payoff response. The tables report these differences alongside their means.

For Assurance, the numeric primary values are $\lambda \in \{0.05, 0.10, 0.20, 0.40\}$. Since $h = 3(1 + \lambda)$, $\beta_h = \beta_\lambda/3$ and $\beta_\lambda(0.35) = \beta_h(1.05)$. Both original- and amended-grid summaries use these coordinate-consistent units and equal game weights. All contrast SDs are formed from the corresponding lineage differences.

Table 11: Endpoint versus all-four-point fitted span drops at HL40 (pp), mean ± seven-lineage SD for Base (B), SFT (S), and Aligned (A). E denotes the observed rate at D_1 minus that at D_4; OLS uses all four numeric coordinates. Each stage has two columns: the observed endpoint drop and the fitted whole-span drop. Mean paired stage contrasts are reported in Table 16.

Game	B: E	B: OLS	S: E	S: OLS	A: E	A: OLS
Assurance	-0.08 ± 0.71	$+0.02 \pm 0.65$	$+5.77 \pm 5.79$	$+7.70 \pm 10.94$	$+8.88 \pm 12.47$	$+11.96 \pm 10.85$
CPR	$+2.41 \pm 5.15$	$+2.59 \pm 4.78$	$+4.26 \pm 7.08$	$+4.62 \pm 7.23$	$+4.12 \pm 4.99$	$+4.19 \pm 4.57$
Donation-PD	$+0.00 \pm 0.63$	-0.23 ± 1.78	$+2.55 \pm 4.66$	$+1.83 \pm 3.99$	$+2.95 \pm 3.26$	$+2.32 \pm 2.73$
PGG	$+2.61 \pm 5.89$	$+2.89 \pm 6.41$	$+9.52 \pm 8.89$	$+9.75 \pm 9.13$	$+10.74 \pm 10.89$	$+10.46 \pm 11.22$
TD	$+1.64 \pm 5.43$	$+1.72 \pm 5.74$	-0.20 ± 1.85	-0.32 ± 1.93	$+2.21 \pm 2.50$	$+2.17 \pm 1.94$
Volunteer	$+10.36 \pm 19.71$	$+9.77 \pm 18.16$	$+9.60 \pm 6.91$	$+10.17 \pm 7.38$	$+12.23 \pm 9.42$	$+12.83 \pm 10.70$

Table 12: Panel history contrasts: SD is across seven lineages. The exploratory 95% t_6 interval assesses model-composition sensitivity; the last two columns count negative and positive lineage estimates. The six-binary-game panel includes Trust with all-seat weights and excludes TD; the six-symmetric-game panel includes TD and excludes Trust.

Contrast	Panel	Mean ± SD	Model interval	Negative	Positive
H_B	Six binary	-1.17 ± 3.92	[-4.79, 2.45]	2/7	5/7
H_S	Six binary	-8.11 ± 5.46	[-13.16, -3.05]	7/7	0/7
H_A	Six binary	-10.93 ± 6.40	[-16.85, -5.01]	7/7	0/7
$K_{S/B}$	Six binary	-6.93 ± 7.30	[-13.68, -0.19]	6/7	1/7
$K_{A/S}$	Six binary	-2.82 ± 4.37	[-6.87, 1.22]	6/7	1/7
$K_{A/B}$	Six binary	-9.76 ± 7.47	[-16.67, -2.85]	6/7	1/7
T_B	Six symmetric	$+1.40 \pm 3.42$	[-1.76, 4.56]	4/7	3/7
T_S	Six symmetric	$+2.47 \pm 2.51$	[0.14, 4.79]	1/7	6/7
T_A	Six symmetric	$+3.74 \pm 2.86$	[1.10, 6.38]	1/7	6/7
$Q_{S/B}$	Six symmetric	$+1.07 \pm 4.09$	[-2.71, 4.85]	2/7	5/7
$Q_{A/B}$	Six symmetric	$+2.34 \pm 4.76$	[-2.07, 6.75]	2/7	5/7

Table 13: History effects by game (pp; mean ± SD across seven lineages). H_s is the stage's HL40-minus-HL2 change; $K_{s/t} = H_s - H_t$ compares that change between stages. Negative K means a larger decline, or smaller increase, at the first-named stage. Aligned has a more negative mean change than SFT and Base in each of the five symmetric binary games.

Game	H_B	H_S	H_A	$K_{S/B}$	$K_{A/S}$	$K_{A/B}$
PGG	-3.42 ± 4.91	-6.63 ± 5.32	-11.27 ± 5.77	-3.20 ± 9.53	-4.65 ± 7.14	-7.85 ± 9.54
CPR	-0.37 ± 9.09	-12.34 ± 15.09	-17.11 ± 10.06	-11.96 ± 12.77	-4.77 ± 8.08	-16.73 ± 12.70
Donation-PD	-0.58 ± 2.16	-7.07 ± 7.92	-10.10 ± 10.11	-6.48 ± 8.12	-3.03 ± 4.80	-9.52 ± 10.67
Volunteer	-2.02 ± 5.01	-15.86 ± 14.97	-16.29 ± 18.41	-13.84 ± 17.87	-0.43 ± 11.66	-14.27 ± 20.10
Assurance	-0.55 ± 3.63	$+0.63 \pm 6.65$	-5.45 ± 9.36	$+1.18 \pm 7.34$	-6.08 ± 11.36	-4.90 ± 11.72
TD	$+2.04 \pm 13.03$	$+1.36 \pm 6.38$	-0.19 ± 9.88	-0.67 ± 15.49	-1.55 ± 4.15	-2.23 ± 17.85
Trust	-0.08 ± 4.54	-7.38 ± 7.83	-5.35 ± 10.04	-7.29 ± 9.77	$+2.03 \pm 6.17$	-5.27 ± 10.25

Table 14: How often the history comparison is negative on the 84-configuration grid. L counts lineages after averaging all configurations of a game. P counts each lineage at each of four main parameter values, averaging other parameters. C keeps every complete configuration separate within each lineage. The larger denominators give more detailed views of the same seven lineages.

Game	$K_{S/B}$ L	P	C	$K_{A/B}$ L	P	C	$K_{A/S}$ L	P	C
PGG	5/7	19/28	39/56	4/7	20/28	42/56	6/7	22/28	41/56
CPR	5/7	20/28	82/112	6/7	23/28	93/112	6/7	20/28	82/112
Donation-PD	5/7	20/28	56/84	5/7	22/28	65/84	4/7	18/28	57/84
Volunteer	6/7	22/28	62/84	5/7	20/28	59/84	4/7	15/28	43/84
Assurance	4/7	13/28	19/56	5/7	19/28	27/56	5/7	20/28	35/56
TD	3/7	12/28	47/112	3/7	12/28	48/112	4/7	14/28	54/112
Trust	6/7	22/28	63/84	5/7	17/28	50/84	3/7	12/28	34/84

In all five symmetric binary-game means, Aligned's history change is more negative than SFT's and Base's ($K_{A/S} < 0$ and $K_{A/B} < 0$). These are five game averages from the same seven lineages. Assurance illustrates the difference between a rate and a change: Aligned cooperates more than SFT at HL40, but loses more cooperation when moving from HL2 to HL40. For SFT relative to Base, the intervals include zero on the original 82 grid and on the five-game summary. The Aligned-relative-to-SFT interval also includes zero. These intervals and direction counts show how the average comparisons vary across models and game sets.

Linear contrasts and probability-scale responses. On a fixed grid, an OLS slope with intercept has weights proportional to $z_j - \bar{z}$, which sum to zero. A constant percentage-point shift therefore changes the level but not the slope. A positive affine rescaling of the coordinate divides the slope and multiplies its span by the same factor, preserving S. By substitution, $K_{S/B} = H_S - H_B = [R_S(40) - R_B(40)] - [R_S(2) - R_B(2)]$; a negative value means that the signed SFT-minus-Base difference is lower at HL40, not necessarily that its absolute magnitude is larger. The corresponding identity holds for Aligned.

A constant change ω in log odds transforms p into $f_\omega(p) = e^\omega p/(1 - p + e^\omega p)$ and can change both its mean and probability-scale payoff differences. For example, $(0.95, 0.80)$ with $\omega = -2$ becomes approximately $(0.720, 0.351)$: the mean falls while the gap grows. Thus S captures changes in observed probability-scale responses, including those produced by shifts in baseline log odds.

G History Effects for Individual Payoff Configurations

The main history figure shows four values of the main payoff parameter for each lineage, averaging runs and other parameters. Figure 3 instead shows each complete configuration separately. Both figures use the same observations and have the same game means.

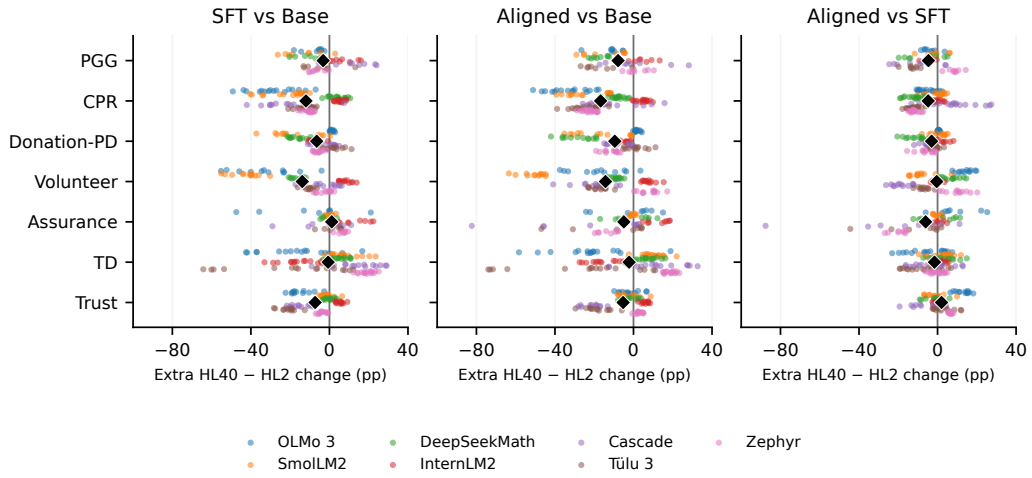

Figure 3: Full-configuration history interactions. Each colored dot is one lineage-by-config contrast after five-seed averaging. In game order there are 56, 112, 84, 84, 56, 112, and 84 dots per panel. The black diamonds show the same game means as Figure 2. The three panels in this figure share a scale. Expanded dots reveal within-lineage heterogeneity across the complete seven-lineage panel.

H Payoff Responses by Model

This section expands the main-text payoff table in two ways. Table 16 adds the mean SFT-minus-Base and Aligned-minus-Base differences. Tables 17–22 show each lineage separately, with SDs across its five runs. Within each run, we average secondary parameters and retain the four main-parameter values. Stage differences match runs by seed ID; the same ID need not produce identical random

draws in separate deployments (Appendix D). For every complete payoff configuration, rather than these four-value summaries, see Table 35. Models with high Base cooperation can also have high SFT

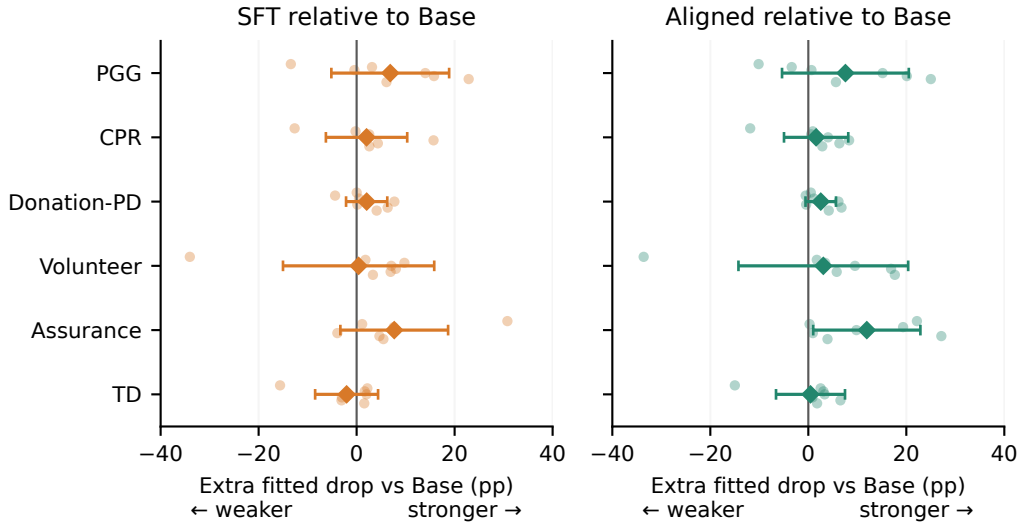

Figure 4: How much larger is the fitted payoff drop after post-training? Panels subtract Base's drop from SFT's (left) or Aligned's (right). Positive values mean that cooperation falls more after post-training across the tested payoff range. Each dot is one lineage's difference, averaged over runs and secondary parameters; diamonds and bars show the mean ± SD across seven lineages. Both panels use the same scale. TD measures normalized claims.

cooperation. Their absolute rates may therefore vary widely while their SFT-minus-Base differences are similar. We compare stages by subtracting within each lineage, as in Figure 4. To distinguish variation across models from variation across payoff settings, let $d_{\ell c}$ be the HL40 SFT-minus-Base difference in pp for lineage ℓ and configuration c in one game. Let $\bar{d}_\ell$ average that difference over configurations. Then

$$\mathrm{Var}_{\ell,c}(d_{\ell c}) = \mathrm{Var}_\ell(\bar{d}_\ell) + E_\ell \,\mathrm{Var}_c(d_{\ell c}).$$

Here E_ℓ averages lineages. We calculate both variance terms with equal weights over the observed estimates. Differences between lineage averages account for 90.8% of the variation in PGG and 89.8% in CPR (Table 15). In these two games, the SFT–Base difference therefore varies much more between models than between payoff settings within a model. Appendix A.2 explains how these SDs differ from uncertainty about a mean.

Table 15: Where does variation in the HL40 SFT–Base difference come from? We separate differences between lineage averages from differences among payoff settings within a lineage. Entries use equal lineage/configuration weights and population variance, in pp^2, calculated from the observed estimates.

Game	Total	Between lineage	Within lineage/config
PGG	1586.55	1439.81	146.74
CPR	829.71	745.27	84.44

Table 16: **Full HL40 payoff surfaces and paired stage differences.** Each game's first row gives four values of its main payoff parameter (D_1–D_4), ordered toward higher cooperation costs, lower benefits, or greater rewards for noncooperation. Entries average seven lineages equally, after averaging seeds and secondary parameters within lineage. Stage rows give cooperation rates (%; normalized claims for TD); difference rows give SFT minus Base and Aligned minus Base (pp). Negative differences mean lower cooperation than Base (lower claims in TD). When a negative difference becomes more negative from D_1 to D_4, the post-trained stage falls further below Base. The final column uses a straight-line fit to all four points (Equation 2). In a stage row, a positive drop means a fitted decline; in a difference row, it is the post-trained drop minus Base's, so a positive value means a larger fitted decline than Base's.

Game / stage	D_1	D_2	D_3	D_4	Fitted drop
PGG: MPCR	0.8	0.6	0.5	0.4	–
Base (B)	59.8	58.6	57.1	57.2	+2.9
SFT (S)	43.2	38.6	35.4	33.7	+9.8
Aligned (A)	42.2	38.7	35.6	31.4	+10.5
SFT − Base	−16.6	−20.0	−21.6	−23.5	+6.9
Aligned − Base	−17.7	−19.9	−21.5	−25.8	+7.6
CPR: θ	0.2	0.4	0.6	0.8	–
Base (B)	68.0	63.7	62.3	65.6	+2.6
SFT (S)	31.4	33.2	30.5	27.1	+4.6
Aligned (A)	23.6	26.2	24.6	19.5	+4.2
SFT − Base	−36.6	−30.5	−31.8	−38.5	+2.0
Aligned − Base	−44.4	−37.5	−37.7	−46.1	+1.6
Donation-PD: R	2.4	2	1.4	1.1	–
Base (B)	54.3	52.9	53.7	54.3	−0.2
SFT (S)	15.9	12.6	13.6	13.4	+1.8
Aligned (A)	11.7	8.9	9.3	8.7	+2.3
SFT − Base	−38.4	−40.3	−40.1	−41.0	+2.1
Aligned − Base	−42.7	−44.1	−44.4	−45.6	+2.5
Volunteer: ρ	0.125	0.375	0.625	0.875	–
Base (B)	80.6	76.6	75.1	70.2	+9.8
SFT (S)	51.9	49.3	44.2	42.3	+10.2
Aligned (A)	46.0	42.6	36.5	33.8	+12.8
SFT − Base	−28.6	−27.2	−30.9	−27.8	+0.4
Aligned − Base	−34.5	−34.0	−38.6	−36.4	+3.1
Assurance: λ	0.4	0.2	0.1	0.05	–
Base (B)	87.2	85.7	86.8	87.3	+0.0
SFT (S)	80.8	78.3	71.5	75.0	+7.7
Aligned (A)	85.2	80.0	70.4	76.4	+12.0
SFT − Base	−6.4	−7.3	−15.4	−12.3	+7.7
Aligned − Base	−2.0	−5.7	−16.5	−10.9	+11.9
TD: τ	0.025	0.05	0.1	0.2	–
Base (B)	39.6	39.8	40.7	37.9	+1.7
SFT (S)	67.2	66.9	67.0	67.4	−0.3
Aligned (A)	62.6	62.1	61.4	60.4	+2.2
SFT − Base	+27.6	+27.1	+26.4	+29.5	−2.0
Aligned − Base	+23.0	+22.4	+20.8	+22.4	+0.4

Table 17: **PGG: HL40 payoff responses for each lineage.** Entries are mean ± SD across five runs, averaging secondary parameters within each run. Cooperation rates are percentages; stage differences and fitted declines are in pp. Stage differences match runs by seed ID. The final column fits all four parameter values within each run before taking the mean and SD.

Lineage / stage	D_1	D_2	D_3	D_4	Fitted drop
PGG: MPCR	0.8	0.6	0.5	0.4	–
OLMo-3-7B					
Base (B)	20.3 ± 23.7	10.1 ± 7.3	3.3 ± 2.6	4.8 ± 1.3	$+17.0 \pm 24.9$
SFT (S)	8.5 ± 2.1	7.6 ± 3.0	6.2 ± 1.5	4.9 ± 0.4	$+3.6 \pm 2.1$
Aligned (A)	16.1 ± 6.1	13.4 ± 3.2	11.7 ± 2.9	9.0 ± 2.5	$+6.8 \pm 4.5$
SFT − Base	-11.8 ± 23.9	-2.4 ± 9.3	$+2.9 \pm 2.6$	$+0.1 \pm 1.5$	-13.4 ± 25.5
Aligned − Base	-4.1 ± 28.5	$+3.3 \pm 8.3$	$+8.4 \pm 3.8$	$+4.2 \pm 2.4$	-10.1 ± 27.8
SmolLM2-1.7B					
Base (B)	99.8 ± 0.1	99.8 ± 0.1	100.0 ± 0.1	99.9 ± 0.1	-0.1 ± 0.2
SFT (S)	40.7 ± 7.0	34.6 ± 3.0	37.8 ± 7.2	37.3 ± 5.9	$+3.1 \pm 1.5$
Aligned (A)	34.1 ± 4.8	34.8 ± 3.2	39.4 ± 6.0	36.3 ± 6.2	-3.5 ± 7.4
SFT − Base	-59.1 ± 6.9	-65.1 ± 2.9	-62.2 ± 7.1	-62.6 ± 5.9	$+3.1 \pm 1.5$
Aligned − Base	-65.8 ± 4.8	-64.9 ± 3.1	-60.6 ± 6.0	-63.6 ± 6.1	-3.4 ± 7.4
DeepSeekMath-7B					
Base (B)	98.6 ± 0.2	98.9 ± 0.2	98.9 ± 0.4	98.8 ± 0.3	-0.2 ± 0.3
SFT (S)	35.6 ± 6.6	34.4 ± 6.4	37.5 ± 9.0	35.4 ± 4.8	-0.7 ± 9.1
Aligned (A)	3.8 ± 0.7	3.8 ± 0.7	4.3 ± 1.7	3.1 ± 0.6	$+0.4 \pm 1.2$
SFT − Base	-63.1 ± 6.6	-64.4 ± 6.4	-61.3 ± 8.9	-63.4 ± 4.9	-0.5 ± 9.3
Aligned − Base	-94.8 ± 0.8	-95.1 ± 0.8	-94.5 ± 1.5	-95.7 ± 0.6	$+0.6 \pm 1.3$
InternLM2-7B					
Base (B)	36.9 ± 2.1	37.2 ± 1.5	35.3 ± 2.1	37.9 ± 2.4	-0.3 ± 2.5
SFT (S)	46.4 ± 2.9	43.3 ± 2.9	33.4 ± 1.3	34.4 ± 2.6	$+13.8 \pm 4.4$
Aligned (A)	54.4 ± 0.9	48.7 ± 1.9	41.0 ± 2.7	40.7 ± 2.8	$+14.9 \pm 2.9$
SFT − Base	$+9.6 \pm 4.9$	$+6.1 \pm 3.9$	-1.9 ± 3.3	-3.5 ± 4.4	$+14.0 \pm 6.3$
Aligned − Base	$+17.6 \pm 2.2$	$+11.5 \pm 1.7$	$+5.8 \pm 4.1$	$+2.8 \pm 3.5$	$+15.2 \pm 1.4$
Nemotron-Cascade-8B					
Base (B)	16.8 ± 19.6	17.7 ± 16.4	14.7 ± 12.0	12.8 ± 2.6	$+4.0 \pm 16.5$
SFT (S)	63.9 ± 14.2	56.3 ± 10.2	45.2 ± 22.1	46.1 ± 9.0	$+19.8 \pm 20.0$
Aligned (A)	50.7 ± 9.7	52.9 ± 9.7	32.4 ± 20.1	28.7 ± 5.1	$+24.1 \pm 10.3$
SFT − Base	$+47.2 \pm 26.9$	$+38.7 \pm 19.3$	$+30.6 \pm 30.7$	$+33.3 \pm 8.6$	$+15.8 \pm 29.9$
Aligned − Base	$+33.9 \pm 24.0$	$+35.2 \pm 23.8$	$+17.8 \pm 29.8$	$+15.9 \pm 4.9$	$+20.1 \pm 20.8$
Tülu-3-8B					
Base (B)	98.5 ± 0.1	98.5 ± 0.5	98.5 ± 0.4	98.4 ± 0.3	$+0.1 \pm 0.2$
SFT (S)	50.6 ± 2.6	37.7 ± 4.8	36.1 ± 2.8	26.2 ± 1.0	$+23.0 \pm 1.5$
Aligned (A)	45.4 ± 6.8	26.9 ± 2.4	29.8 ± 5.6	18.0 ± 1.7	$+25.1 \pm 8.6$
SFT − Base	-48.0 ± 2.7	-60.7 ± 5.1	-62.4 ± 3.0	-72.2 ± 1.1	$+22.9 \pm 1.5$
Aligned − Base	-53.1 ± 6.7	-71.6 ± 2.9	-68.7 ± 6.0	-80.4 ± 1.9	$+25.0 \pm 8.8$
Zephyr-7B					
Base (B)	47.9 ± 1.5	47.9 ± 2.1	48.8 ± 1.9	47.9 ± 0.9	-0.3 ± 1.4
SFT (S)	56.8 ± 1.5	55.9 ± 3.3	51.8 ± 1.2	51.6 ± 1.4	$+5.8 \pm 1.4$
Aligned (A)	90.6 ± 2.9	90.5 ± 2.3	90.2 ± 1.2	84.1 ± 1.9	$+5.3 \pm 2.9$
SFT − Base	$+8.9 \pm 0.7$	$+8.0 \pm 2.7$	$+2.9 \pm 2.0$	$+3.8 \pm 1.2$	$+6.1 \pm 0.9$
Aligned − Base	$+42.7 \pm 2.6$	$+42.6 \pm 3.6$	$+41.3 \pm 0.9$	$+36.2 \pm 2.6$	$+5.6 \pm 2.2$

Table 18: **CPR: HL40 payoff responses for each lineage.** Entries are mean ± SD across five runs, averaging secondary parameters within each run. Cooperation rates are percentages; stage differences and fitted declines are in pp. Stage differences match runs by seed ID. The final column fits all four parameter values within each run before taking the mean and SD.

Lineage / stage	D_1	D_2	D_3	D_4	Fitted drop
CPR: θ	0.2	0.4	0.6	0.8	–
OLMo-3-7B					
Base (B)	56.5 ± 8.7	45.9 ± 7.1	45.5 ± 8.4	42.6 ± 9.4	$+12.7 \pm 8.8$
SFT (S)	12.3 ± 2.5	12.6 ± 3.7	13.0 ± 6.3	12.1 ± 2.8	$+0.0 \pm 2.8$
Aligned (A)	9.8 ± 1.6	7.9 ± 0.7	8.2 ± 0.3	8.8 ± 1.5	$+0.8 \pm 1.9$
SFT − Base	-44.2 ± 6.6	-33.3 ± 4.4	-32.5 ± 10.6	-30.4 ± 10.8	-12.7 ± 11.6
Aligned − Base	-46.7 ± 8.2	-38.0 ± 6.5	-37.3 ± 8.3	-33.8 ± 10.7	-11.9 ± 8.5
SmolLM2-1.7B					
Base (B)	99.6 ± 0.1	99.6 ± 0.2	99.6 ± 0.2	99.7 ± 0.1	-0.1 ± 0.1
SFT (S)	8.5 ± 1.5	19.0 ± 2.7	17.1 ± 3.2	9.5 ± 1.4	-0.3 ± 0.9
Aligned (A)	8.4 ± 0.9	12.9 ± 2.7	12.6 ± 1.4	7.6 ± 0.8	$+0.8 \pm 1.5$
SFT − Base	-91.0 ± 1.5	-80.6 ± 2.6	-82.4 ± 3.2	-90.2 ± 1.3	-0.2 ± 1.0
Aligned − Base	-91.2 ± 0.8	-86.6 ± 2.7	-87.0 ± 1.3	-92.1 ± 0.7	$+0.9 \pm 1.5$
DeepSeekMath-7B					
Base (B)	97.8 ± 0.4	83.4 ± 3.2	83.7 ± 1.9	97.8 ± 0.4	-0.0 ± 0.9
SFT (S)	54.1 ± 3.7	66.3 ± 3.0	63.3 ± 3.1	52.3 ± 2.8	$+2.5 \pm 5.5$
Aligned (A)	8.8 ± 1.1	36.5 ± 2.8	34.4 ± 4.2	8.9 ± 0.7	$+0.5 \pm 0.8$
SFT − Base	-43.8 ± 3.9	-17.1 ± 2.0	-20.4 ± 3.5	-45.5 ± 3.1	$+2.6 \pm 5.7$
Aligned − Base	-89.0 ± 0.8	-46.9 ± 4.0	-49.3 ± 4.2	-88.9 ± 0.8	$+0.6 \pm 1.2$
InternLM2-7B					
Base (B)	45.2 ± 1.7	43.6 ± 1.6	42.3 ± 1.3	45.5 ± 1.0	$+0.1 \pm 1.0$
SFT (S)	27.9 ± 2.5	25.0 ± 1.0	24.7 ± 1.0	25.8 ± 1.0	$+2.0 \pm 2.2$
Aligned (A)	40.3 ± 2.0	38.8 ± 1.5	36.5 ± 1.7	36.5 ± 2.0	$+4.1 \pm 3.7$
SFT − Base	-17.3 ± 2.9	-18.5 ± 1.7	-17.7 ± 1.5	-19.7 ± 1.9	$+1.9 \pm 2.0$
Aligned − Base	-4.9 ± 1.7	-4.8 ± 2.1	-5.9 ± 1.9	-9.1 ± 2.8	$+4.0 \pm 3.0$
Nemotron-Cascade-8B					
Base (B)	22.0 ± 10.4	35.3 ± 8.8	25.8 ± 4.9	19.7 ± 2.3	$+4.8 \pm 8.2$
SFT (S)	46.4 ± 7.0	39.4 ± 7.4	30.4 ± 9.1	26.6 ± 5.6	$+20.5 \pm 8.1$
Aligned (A)	32.1 ± 15.4	22.3 ± 2.3	21.4 ± 3.0	17.8 ± 3.5	$+13.2 \pm 15.4$
SFT − Base	$+24.4 \pm 15.9$	$+4.2 \pm 11.2$	$+4.6 \pm 11.6$	$+6.8 \pm 5.5$	$+15.7 \pm 15.2$
Aligned − Base	$+10.2 \pm 18.4$	-13.0 ± 10.2	-4.4 ± 2.9	-1.9 ± 4.3	$+8.3 \pm 17.4$
Tülu-3-8B					
Base (B)	91.6 ± 3.0	75.3 ± 2.9	76.7 ± 2.1	90.5 ± 4.7	$+0.5 \pm 2.1$
SFT (S)	34.5 ± 1.3	34.3 ± 2.0	31.9 ± 1.3	29.9 ± 1.7	$+4.9 \pm 2.0$
Aligned (A)	33.3 ± 1.2	32.6 ± 0.9	28.4 ± 1.0	27.1 ± 0.4	$+6.9 \pm 1.5$
SFT − Base	-57.1 ± 2.4	-41.0 ± 2.8	-44.9 ± 1.9	-60.6 ± 5.7	$+4.3 \pm 4.0$
Aligned − Base	-58.2 ± 2.9	-42.7 ± 2.7	-48.4 ± 2.1	-63.4 ± 4.8	$+6.3 \pm 2.6$
Zephyr-7B					
Base (B)	63.3 ± 3.0	62.8 ± 1.8	62.5 ± 1.3	63.2 ± 2.5	$+0.1 \pm 3.6$
SFT (S)	36.0 ± 1.0	35.5 ± 2.3	33.4 ± 2.2	33.7 ± 2.5	$+2.7 \pm 1.9$
Aligned (A)	32.6 ± 1.2	32.3 ± 1.5	30.7 ± 1.2	29.8 ± 1.6	$+3.0 \pm 1.2$
SFT − Base	-27.2 ± 2.9	-27.3 ± 2.1	-29.0 ± 1.9	-29.6 ± 2.9	$+2.6 \pm 2.5$
Aligned − Base	-30.7 ± 2.9	-30.6 ± 3.1	-31.8 ± 1.7	-33.5 ± 2.5	$+2.8 \pm 4.6$

Table 19: **Donation-PD: HL40 payoff responses for each lineage.** Entries are mean ± SD across five runs, averaging secondary parameters within each run. Cooperation rates are percentages; stage differences and fitted declines are in pp. Stage differences match runs by seed ID. The final column fits all four parameter values within each run before taking the mean and SD.

Lineage / stage	D_1	D_2	D_3	D_4	Fitted drop
Donation-PD: R	2.4	2	1.4	1.1	–
OLMo-3-7B					
Base (B)	2.2 ± 0.3	1.7 ± 0.5	1.4 ± 0.3	1.4 ± 0.3	$+0.8 \pm 0.4$
SFT (S)	1.0 ± 0.2	0.4 ± 0.2	0.2 ± 0.1	0.1 ± 0.0	$+0.8 \pm 0.2$
Aligned (A)	1.7 ± 0.1	1.2 ± 0.2	0.6 ± 0.1	0.4 ± 0.1	$+1.3 \pm 0.1$
SFT − Base	-1.3 ± 0.4	-1.3 ± 0.6	-1.2 ± 0.3	-1.3 ± 0.3	$+0.0 \pm 0.5$
Aligned − Base	-0.5 ± 0.2	-0.5 ± 0.5	-0.8 ± 0.4	-1.0 ± 0.4	$+0.5 \pm 0.4$
SmolLM2-1.7B					
Base (B)	99.7 ± 0.1	99.8 ± 0.1	99.7 ± 0.1	99.7 ± 0.3	$+0.1 \pm 0.2$
SFT (S)	7.8 ± 3.4	9.7 ± 2.4	11.7 ± 4.7	12.1 ± 4.2	-4.3 ± 4.9
Aligned (A)	3.4 ± 0.8	5.6 ± 1.7	5.6 ± 1.8	3.8 ± 1.0	-0.5 ± 2.0
SFT − Base	-91.9 ± 3.4	-90.1 ± 2.4	-88.0 ± 4.7	-87.6 ± 4.2	-4.4 ± 4.8
Aligned − Base	-96.3 ± 0.8	-94.2 ± 1.6	-94.1 ± 1.8	-95.8 ± 1.2	-0.5 ± 2.1
DeepSeekMath-7B					
Base (B)	95.9 ± 0.2	96.1 ± 0.6	95.9 ± 0.5	96.0 ± 0.7	-0.0 ± 0.8
SFT (S)	12.0 ± 2.2	5.8 ± 0.9	9.0 ± 1.6	10.2 ± 1.7	$+0.5 \pm 1.1$
Aligned (A)	5.8 ± 0.7	1.8 ± 0.7	3.5 ± 1.0	3.7 ± 0.7	$+1.1 \pm 0.6$
SFT − Base	-83.9 ± 2.2	-90.3 ± 1.1	-86.9 ± 1.9	-85.8 ± 1.6	$+0.5 \pm 1.4$
Aligned − Base	-90.2 ± 0.5	-94.2 ± 1.0	-92.4 ± 1.1	-92.3 ± 1.2	$+1.1 \pm 1.1$
InternLM2-7B					
Base (B)	43.0 ± 1.8	41.7 ± 1.2	41.0 ± 0.6	42.7 ± 1.5	$+0.6 \pm 1.8$
SFT (S)	37.7 ± 1.4	28.7 ± 1.1	29.9 ± 1.8	27.3 ± 1.1	$+8.3 \pm 0.6$
Aligned (A)	41.1 ± 2.5	31.8 ± 2.1	32.7 ± 1.6	32.9 ± 1.8	$+6.8 \pm 2.6$
SFT − Base	-5.3 ± 2.1	-13.0 ± 1.3	-11.1 ± 1.3	-15.4 ± 2.6	$+7.7 \pm 2.1$
Aligned − Base	-1.9 ± 3.0	-9.9 ± 3.3	-8.3 ± 1.4	-9.9 ± 2.5	$+6.2 \pm 3.1$
Nemotron-Cascade-8B					
Base (B)	1.0 ± 0.3	0.4 ± 0.1	0.7 ± 0.3	0.7 ± 0.2	$+0.2 \pm 0.1$
SFT (S)	5.5 ± 1.7	3.4 ± 0.4	3.8 ± 0.5	5.1 ± 1.8	$+0.3 \pm 0.8$
Aligned (A)	3.3 ± 1.5	1.6 ± 0.3	3.3 ± 0.9	2.9 ± 0.5	-0.3 ± 1.3
SFT − Base	$+4.5 \pm 1.5$	$+3.0 \pm 0.4$	$+3.1 \pm 0.7$	$+4.4 \pm 1.6$	$+0.1 \pm 0.7$
Aligned − Base	$+2.3 \pm 1.4$	$+1.2 \pm 0.4$	$+2.6 \pm 0.9$	$+2.2 \pm 0.4$	-0.4 ± 1.4
Tülu-3-8B					
Base (B)	91.4 ± 4.6	81.1 ± 8.6	90.4 ± 6.4	92.6 ± 4.2	-4.2 ± 4.5
SFT (S)	15.8 ± 1.0	14.2 ± 0.6	13.7 ± 1.4	13.5 ± 0.7	$+2.2 \pm 0.7$
Aligned (A)	7.3 ± 1.5	6.5 ± 1.3	4.8 ± 0.9	4.9 ± 1.5	$+2.6 \pm 1.2$
SFT − Base	-75.5 ± 4.7	-66.9 ± 8.8	-76.7 ± 6.7	-79.1 ± 4.3	$+6.3 \pm 4.4$
Aligned − Base	-84.1 ± 3.4	-74.5 ± 8.4	-85.6 ± 6.0	-87.7 ± 4.1	$+6.8 \pm 4.7$
Zephyr-7B					
Base (B)	47.1 ± 1.5	50.0 ± 2.3	46.9 ± 1.4	47.2 ± 2.1	$+1.0 \pm 2.5$
SFT (S)	31.6 ± 1.8	26.1 ± 1.0	26.7 ± 2.1	25.3 ± 1.5	$+5.0 \pm 1.3$
Aligned (A)	19.0 ± 3.4	13.7 ± 1.5	14.9 ± 0.9	12.3 ± 2.1	$+5.1 \pm 3.2$
SFT − Base	-15.5 ± 1.6	-23.8 ± 2.4	-20.2 ± 3.2	-21.9 ± 3.1	$+4.1 \pm 3.2$
Aligned − Base	-28.1 ± 4.6	-36.2 ± 2.4	-32.0 ± 1.2	-35.0 ± 2.0	$+4.2 \pm 3.8$

Table 20: **Volunteer: HL40 payoff responses for each lineage.** Entries are mean ± SD across five runs, averaging secondary parameters within each run. Cooperation rates are percentages; stage differences and fitted declines are in pp. Stage differences match runs by seed ID. The final column fits all four parameter values within each run before taking the mean and SD.

Lineage / stage	D_1	D_2	D_3	D_4	Fitted drop
Volunteer: ρ	0.125	0.375	0.625	0.875	–
OLMo-3-7B					
Base (B)	83.4 ± 3.6	59.5 ± 15.2	56.2 ± 13.8	30.3 ± 3.7	$+48.8 \pm 7.4$
SFT (S)	41.3 ± 2.2	35.8 ± 0.7	29.4 ± 0.8	27.0 ± 1.6	$+14.8 \pm 2.0$
Aligned (A)	45.2 ± 2.4	39.3 ± 2.5	34.6 ± 3.6	29.8 ± 2.2	$+15.2 \pm 2.4$
SFT − Base	-42.1 ± 4.3	-23.7 ± 15.3	-26.8 ± 14.1	-3.2 ± 4.2	-34.0 ± 5.7
Aligned − Base	-38.3 ± 2.9	-20.2 ± 14.4	-21.6 ± 10.9	-0.5 ± 4.0	-33.6 ± 5.8
SmolLM2-1.7B					
Base (B)	99.6 ± 0.2	99.6 ± 0.1	99.6 ± 0.1	99.5 ± 0.1	$+0.0 \pm 0.2$
SFT (S)	38.8 ± 6.4	37.9 ± 4.2	36.2 ± 6.8	37.3 ± 3.3	$+1.8 \pm 4.5$
Aligned (A)	32.7 ± 4.7	28.5 ± 2.9	29.3 ± 8.0	30.5 ± 6.8	$+1.8 \pm 10.7$
SFT − Base	-60.8 ± 6.3	-61.7 ± 4.3	-63.5 ± 6.8	-62.2 ± 3.2	$+1.8 \pm 4.5$
Aligned − Base	-66.9 ± 4.6	-71.1 ± 2.9	-70.3 ± 8.0	-69.1 ± 6.8	$+1.7 \pm 10.6$
DeepSeekMath-7B					
Base (B)	97.2 ± 0.4	97.3 ± 0.2	96.9 ± 0.4	96.2 ± 0.2	$+1.0 \pm 0.4$
SFT (S)	45.9 ± 2.5	43.8 ± 3.4	38.4 ± 2.1	35.8 ± 1.6	$+10.7 \pm 1.7$
Aligned (A)	16.5 ± 1.0	14.9 ± 0.9	13.7 ± 0.4	12.0 ± 1.0	$+4.4 \pm 1.1$
SFT − Base	-51.3 ± 2.5	-53.5 ± 3.4	-58.5 ± 1.9	-60.4 ± 1.8	$+9.7 \pm 1.8$
Aligned − Base	-80.7 ± 0.9	-82.4 ± 1.0	-83.1 ± 0.8	-84.2 ± 1.0	$+3.4 \pm 0.7$
InternLM2-7B					
Base (B)	35.5 ± 0.8	36.0 ± 1.3	36.0 ± 2.2	36.1 ± 1.4	-0.5 ± 1.7
SFT (S)	45.1 ± 0.9	43.5 ± 0.9	41.1 ± 1.4	38.6 ± 1.6	$+6.6 \pm 2.3$
Aligned (A)	53.8 ± 2.2	50.0 ± 1.2	45.7 ± 1.0	45.2 ± 1.6	$+9.0 \pm 2.3$
SFT − Base	$+9.6 \pm 0.9$	$+7.5 \pm 1.8$	$+5.2 \pm 3.4$	$+2.5 \pm 1.9$	$+7.1 \pm 2.4$
Aligned − Base	$+18.3 \pm 2.7$	$+13.9 \pm 1.5$	$+9.7 \pm 1.8$	$+9.1 \pm 2.9$	$+9.5 \pm 3.7$
Nemotron-Cascade-8B					
Base (B)	98.3 ± 1.7	94.5 ± 4.1	89.1 ± 8.0	82.3 ± 8.6	$+16.0 \pm 8.5$
SFT (S)	86.5 ± 3.6	81.4 ± 4.8	68.1 ± 2.3	64.3 ± 7.2	$+24.0 \pm 7.4$
Aligned (A)	72.2 ± 6.6	69.2 ± 5.7	48.0 ± 2.9	42.8 ± 1.7	$+32.9 \pm 6.9$
SFT − Base	-11.7 ± 3.5	-13.1 ± 7.7	-21.0 ± 8.3	-18.0 ± 7.3	$+8.0 \pm 7.8$
Aligned − Base	-26.0 ± 7.9	-25.3 ± 6.3	-41.0 ± 7.4	-39.5 ± 7.3	$+16.9 \pm 12.9$
Tülu-3-8B					
Base (B)	98.6 ± 0.4	98.0 ± 0.5	97.5 ± 0.1	97.1 ± 0.5	$+1.5 \pm 0.8$
SFT (S)	45.9 ± 1.7	44.1 ± 1.2	40.8 ± 1.5	37.6 ± 0.8	$+8.4 \pm 1.7$
Aligned (A)	37.9 ± 1.5	35.0 ± 1.5	33.6 ± 2.3	30.3 ± 1.1	$+7.3 \pm 1.5$
SFT − Base	-52.7 ± 1.9	-54.0 ± 1.3	-56.8 ± 1.5	-59.5 ± 0.6	$+6.9 \pm 1.7$
Aligned − Base	-60.7 ± 1.6	-63.0 ± 1.9	-63.9 ± 2.4	-66.8 ± 1.5	$+5.8 \pm 1.9$
Zephyr-7B					
Base (B)	51.4 ± 1.8	51.0 ± 0.9	50.5 ± 1.7	49.8 ± 1.6	$+1.5 \pm 1.2$
SFT (S)	60.2 ± 0.9	58.8 ± 2.1	55.7 ± 1.4	55.9 ± 1.5	$+4.9 \pm 2.1$
Aligned (A)	63.9 ± 5.0	60.9 ± 3.1	50.3 ± 1.9	46.1 ± 2.3	$+19.2 \pm 6.0$
SFT − Base	$+8.9 \pm 2.0$	$+7.8 \pm 2.2$	$+5.1 \pm 2.6$	$+6.0 \pm 2.3$	$+3.3 \pm 2.3$
Aligned − Base	$+12.6 \pm 5.8$	$+9.9 \pm 3.4$	-0.2 ± 2.1	-3.7 ± 1.4	$+17.7 \pm 5.8$

Table 21: **Assurance: HL40 payoff responses for each lineage.** Entries are mean ± SD across five runs, averaging secondary parameters within each run. Cooperation rates are percentages; stage differences and fitted declines are in pp. Stage differences match runs by seed ID. The final column fits all four parameter values within each run before taking the mean and SD.

Lineage / stage	D_1	D_2	D_3	D_4	Fitted drop
Assurance: λ	0.4	0.2	0.1	0.05	–
OLMo-3-7B					
Base (B)	98.3 ± 0.6	89.8 ± 19.5	98.0 ± 0.2	98.0 ± 0.5	-0.2 ± 1.0
SFT (S)	89.9 ± 18.2	90.3 ± 13.0	45.6 ± 16.5	72.4 ± 20.3	$+30.6 \pm 16.1$
Aligned (A)	81.8 ± 19.6	71.3 ± 22.7	57.0 ± 4.4	63.8 ± 18.8	$+22.0 \pm 30.0$
SFT − Base	-8.5 ± 18.2	$+0.5 \pm 6.9$	-52.5 ± 16.5	-25.6 ± 20.4	$+30.8 \pm 16.4$
Aligned − Base	-16.6 ± 20.1	-18.5 ± 37.9	-41.0 ± 4.5	-34.2 ± 18.6	$+22.2 \pm 30.7$
SmolLM2-1.7B					
Base (B)	99.1 ± 0.3	99.1 ± 0.4	99.3 ± 0.4	99.5 ± 0.4	-0.3 ± 0.2
SFT (S)	96.8 ± 3.6	93.8 ± 6.5	95.2 ± 5.0	96.2 ± 3.6	$+0.8 \pm 3.2$
Aligned (A)	97.6 ± 4.0	95.8 ± 4.6	96.7 ± 3.7	98.0 ± 3.2	-0.0 ± 0.9
SFT − Base	-2.4 ± 3.5	-5.3 ± 6.1	-4.1 ± 4.7	-3.2 ± 3.2	$+1.1 \pm 3.3$
Aligned − Base	-1.6 ± 3.8	-3.3 ± 4.1	-2.6 ± 3.4	-1.4 ± 2.9	$+0.3 \pm 1.0$
DeepSeekMath-7B					
Base (B)	99.2 ± 0.3	99.2 ± 0.2	99.0 ± 0.3	99.2 ± 0.2	$+0.1 \pm 0.3$
SFT (S)	87.3 ± 2.3	83.1 ± 3.2	77.0 ± 3.6	82.1 ± 3.1	$+7.7 \pm 3.1$
Aligned (A)	87.7 ± 8.2	78.0 ± 9.5	61.0 ± 11.3	74.7 ± 18.1	$+19.4 \pm 5.1$
SFT − Base	-11.9 ± 2.3	-16.1 ± 3.3	-22.1 ± 3.6	-17.1 ± 3.2	$+7.6 \pm 3.2$
Aligned − Base	-11.6 ± 8.2	-21.1 ± 9.6	-38.0 ± 11.4	-24.5 ± 18.0	$+19.3 \pm 4.8$
InternLM2-7B					
Base (B)	40.6 ± 2.1	41.6 ± 1.6	40.2 ± 1.9	42.0 ± 1.5	-0.7 ± 3.4
SFT (S)	71.1 ± 2.8	69.1 ± 2.2	64.3 ± 2.9	64.3 ± 1.1	$+7.3 \pm 2.5$
Aligned (A)	84.5 ± 2.4	82.2 ± 1.8	78.0 ± 4.8	74.8 ± 3.3	$+9.1 \pm 4.4$
SFT − Base	$+30.5 \pm 2.9$	$+27.5 \pm 3.3$	$+24.0 \pm 3.0$	$+22.3 \pm 2.2$	$+8.1 \pm 4.0$
Aligned − Base	$+43.9 \pm 4.4$	$+40.7 \pm 0.9$	$+37.8 \pm 6.2$	$+32.8 \pm 3.8$	$+9.8 \pm 7.5$
Nemotron-Cascade-8B					
Base (B)	99.8 ± 0.1	99.7 ± 0.2	99.8 ± 0.0	99.8 ± 0.1	-0.1 ± 0.2
SFT (S)	59.3 ± 18.3	52.1 ± 1.4	67.7 ± 21.7	59.3 ± 35.3	-4.0 ± 21.9
Aligned (A)	70.8 ± 25.3	71.8 ± 24.4	53.5 ± 2.3	80.7 ± 23.4	$+0.9 \pm 20.2$
SFT − Base	-40.5 ± 18.2	-47.6 ± 1.4	-32.2 ± 21.7	-40.5 ± 35.3	-3.9 ± 21.8
Aligned − Base	-28.9 ± 25.3	-27.9 ± 24.4	-46.3 ± 2.3	-19.2 ± 23.5	$+0.9 \pm 20.1$
Tülu-3-8B					
Base (B)	99.0 ± 0.1	98.9 ± 0.3	99.1 ± 0.2	99.1 ± 0.2	-0.0 ± 0.2
SFT (S)	81.2 ± 6.2	83.9 ± 2.0	77.3 ± 3.9	77.0 ± 5.8	$+4.6 \pm 5.4$
Aligned (A)	88.0 ± 15.9	80.7 ± 19.5	67.5 ± 15.4	60.0 ± 2.6	$+27.1 \pm 15.1$
SFT − Base	-17.9 ± 6.3	-15.0 ± 2.0	-21.8 ± 3.8	-22.1 ± 5.8	$+4.7 \pm 5.3$
Aligned − Base	-11.0 ± 15.9	-18.2 ± 19.6	-31.6 ± 15.4	-39.1 ± 2.4	$+27.1 \pm 15.1$
Zephyr-7B					
Base (B)	74.5 ± 1.6	71.5 ± 2.8	72.3 ± 1.4	73.6 ± 2.3	$+1.4 \pm 3.0$
SFT (S)	80.0 ± 2.5	76.0 ± 1.4	73.2 ± 1.6	73.8 ± 2.8	$+6.8 \pm 3.2$
Aligned (A)	86.4 ± 3.0	80.2 ± 4.8	78.9 ± 3.1	82.7 ± 5.3	$+5.3 \pm 1.2$
SFT − Base	$+5.5 \pm 2.4$	$+4.6 \pm 3.7$	$+0.9 \pm 3.0$	$+0.2 \pm 3.1$	$+5.5 \pm 3.7$
Aligned − Base	$+11.8 \pm 4.4$	$+8.8 \pm 5.0$	$+6.6 \pm 3.3$	$+9.1 \pm 3.3$	$+3.9 \pm 3.1$

Table 22: **TD: HL40 payoff responses for each lineage.** Entries are mean ± SD across five runs, averaging secondary parameters within each run. Normalized claims are multiplied by 100; stage differences and fitted declines are in pp. Stage differences match runs by seed ID. The final column fits all four parameter values within each run before taking the mean and SD.

Lineage / stage	D_1	D_2	D_3	D_4	Fitted drop
TD: τ	0.025	0.05	0.1	0.2	–
OLMo-3-7B					
Base (B)	90.1 ± 5.2	90.9 ± 2.9	88.5 ± 3.1	76.4 ± 10.5	$+14.6 \pm 10.0$
SFT (S)	98.1 ± 1.4	96.9 ± 5.3	98.4 ± 2.1	98.6 ± 1.0	-1.0 ± 2.6
Aligned (A)	83.1 ± 4.1	83.6 ± 9.6	84.1 ± 10.2	83.7 ± 8.9	-0.4 ± 7.4
SFT − Base	$+8.1 \pm 6.2$	$+6.0 \pm 6.6$	$+9.8 \pm 4.1$	$+22.2 \pm 10.2$	-15.6 ± 11.0
Aligned − Base	-7.0 ± 8.3	-7.3 ± 11.3	-4.5 ± 12.8	$+7.2 \pm 19.0$	-15.0 ± 15.4
SmolLM2-1.7B					
Base (B)	3.6 ± 0.4	4.1 ± 1.2	4.4 ± 0.4	3.9 ± 0.4	-0.1 ± 0.6
SFT (S)	62.4 ± 4.6	62.6 ± 5.9	62.6 ± 4.7	60.5 ± 9.2	$+2.0 \pm 9.2$
Aligned (A)	68.3 ± 6.1	67.6 ± 5.6	65.3 ± 4.0	65.9 ± 1.5	$+2.4 \pm 2.5$
SFT − Base	$+58.8 \pm 4.4$	$+58.4 \pm 6.4$	$+58.2 \pm 5.0$	$+56.5 \pm 9.0$	$+2.2 \pm 9.2$
Aligned − Base	$+64.6 \pm 6.4$	$+63.4 \pm 6.3$	$+60.9 \pm 4.3$	$+61.9 \pm 1.7$	$+2.5 \pm 3.0$
DeepSeekMath-7B					
Base (B)	30.5 ± 2.1	32.1 ± 3.8	33.2 ± 3.3	32.2 ± 2.7	-1.2 ± 2.8
SFT (S)	69.0 ± 3.9	68.3 ± 1.3	69.8 ± 3.6	68.2 ± 1.2	$+0.4 \pm 3.7$
Aligned (A)	68.5 ± 2.4	71.1 ± 5.1	70.0 ± 4.0	67.7 ± 5.9	$+1.9 \pm 5.0$
SFT − Base	$+38.5 \pm 3.2$	$+36.2 \pm 4.1$	$+36.7 \pm 4.3$	$+36.0 \pm 3.7$	$+1.7 \pm 5.1$
Aligned − Base	$+38.0 \pm 2.1$	$+39.0 \pm 6.6$	$+36.8 \pm 4.0$	$+35.5 \pm 5.0$	$+3.1 \pm 4.6$
InternLM2-7B					
Base (B)	36.2 ± 2.7	35.6 ± 3.1	36.8 ± 2.2	37.4 ± 1.2	-1.5 ± 2.1
SFT (S)	63.6 ± 0.9	63.0 ± 1.4	63.2 ± 0.7	62.9 ± 1.0	$+0.5 \pm 0.9$
Aligned (A)	63.1 ± 0.9	62.1 ± 0.9	62.2 ± 1.8	61.0 ± 0.8	$+1.8 \pm 1.2$
SFT − Base	$+27.5 \pm 3.3$	$+27.4 \pm 2.1$	$+26.5 \pm 2.5$	$+25.6 \pm 1.4$	$+2.0 \pm 2.5$
Aligned − Base	$+27.0 \pm 3.4$	$+26.6 \pm 3.3$	$+25.5 \pm 3.8$	$+23.7 \pm 1.3$	$+3.3 \pm 3.2$
Nemotron-Cascade-8B					
Base (B)	58.5 ± 7.3	57.0 ± 5.2	60.7 ± 3.2	56.6 ± 13.0	$+1.3 \pm 13.9$
SFT (S)	23.0 ± 4.6	25.2 ± 5.6	25.0 ± 4.7	25.5 ± 2.5	-1.7 ± 2.6
Aligned (A)	25.4 ± 3.8	26.4 ± 3.3	27.1 ± 6.5	23.6 ± 2.8	$+2.1 \pm 3.1$
SFT − Base	-35.5 ± 6.9	-31.8 ± 10.3	-35.6 ± 3.4	-31.1 ± 14.0	-3.0 ± 13.0
Aligned − Base	-33.2 ± 10.3	-30.6 ± 3.6	-33.6 ± 7.5	-32.9 ± 14.8	$+0.8 \pm 16.5$
Tülu-3-8B					
Base (B)	26.6 ± 5.9	27.3 ± 3.5	28.7 ± 2.6	27.2 ± 3.2	-0.5 ± 3.4
SFT (S)	83.3 ± 3.6	81.4 ± 2.8	79.9 ± 8.1	86.2 ± 4.6	-3.6 ± 5.6
Aligned (A)	75.9 ± 7.2	71.1 ± 5.6	69.6 ± 11.7	68.4 ± 7.9	$+6.0 \pm 13.3$
SFT − Base	$+56.7 \pm 4.9$	$+54.1 \pm 5.1$	$+51.2 \pm 9.4$	$+58.9 \pm 4.2$	-3.1 ± 3.1
Aligned − Base	$+49.3 \pm 3.5$	$+43.8 \pm 5.5$	$+40.9 \pm 11.8$	$+41.2 \pm 9.2$	$+6.6 \pm 11.2$
Zephyr-7B					
Base (B)	31.6 ± 3.1	31.4 ± 1.7	32.4 ± 1.3	31.9 ± 2.8	-0.4 ± 2.7
SFT (S)	71.1 ± 1.7	71.1 ± 1.4	70.1 ± 1.8	70.0 ± 2.4	$+1.2 \pm 1.8$
Aligned (A)	53.9 ± 1.4	53.1 ± 0.8	51.8 ± 2.5	52.3 ± 2.2	$+1.4 \pm 2.3$
SFT − Base	$+39.5 \pm 4.2$	$+39.6 \pm 2.7$	$+37.7 \pm 1.9$	$+38.1 \pm 3.3$	$+1.6 \pm 3.5$
Aligned − Base	$+22.3 \pm 4.2$	$+21.7 \pm 2.2$	$+19.4 \pm 2.6$	$+20.5 \pm 2.6$	$+1.8 \pm 2.6$

I Investment and Return Intentions in Trust

All three players decide from the same completed history, before seeing anyone's current action. We ask both trustees whether they would return funds even when the investor chooses to keep its endowment. The trustee cooperation rate therefore measures *return intentions*: the fraction of queried decisions that choose to return. A transfer actually occurs only if the investor invests. Table 23 separates the roles and shows how the weighting changes their combined score.

Table 23: Trust intentions and weighting: mean ± SD across seven lineage contrasts. All-seat weights investor/trustee rates 1/3 and 2/3; role-equal uses 1/2 each. Trustee intentions are elicited simultaneously with the investment decision.

Metric	Exposure	Investor	Trustee	All-seat	Role-equal
$D_{S/B}$	one-shot	-14.11 ± 29.30	-17.11 ± 19.64	-16.11 ± 17.60	-15.61 ± 18.96
$D_{S/B}$	HL2	-14.07 ± 19.96	-21.59 ± 17.41	-19.08 ± 14.60	-17.83 ± 14.63
$D_{S/B}$	HL40	-18.30 ± 25.23	-30.42 ± 27.29	-26.38 ± 23.36	-24.36 ± 22.52
$D_{A/S}$	HL40	$+1.36 \pm 19.73$	-0.36 ± 12.65	$+0.21 \pm 12.99$	$+0.50 \pm 14.09$
H_B	HL40-HL2	$+1.61 \pm 6.14$	-0.93 ± 4.08	-0.08 ± 4.54	$+0.34 \pm 4.87$
H_S	HL40-HL2	-2.62 ± 7.43	-9.76 ± 8.85	-7.38 ± 7.83	-6.19 ± 7.50
H_A	HL40-HL2	-0.59 ± 13.54	-7.73 ± 12.24	-5.35 ± 10.04	-4.16 ± 9.94
$K_{S/B}$	HL40-HL2	-4.23 ± 9.40	-8.83 ± 11.04	-7.29 ± 9.77	-6.53 ± 9.37
$K_{A/S}$	HL40-HL2	$+2.04 \pm 6.67$	$+2.02 \pm 9.86$	$+2.03 \pm 6.17$	$+2.03 \pm 4.91$
$K_{A/B}$	HL40-HL2	-2.20 ± 14.11	-6.80 ± 13.06	-5.27 ± 10.25	-4.50 ± 10.02

J Exploratory Interpretation: Thresholds and Older Records

Table 24: Does the Assurance participation threshold change the history effect? $k = 2$ requires at least two contributors; $k = 3$ requires all three. Entries are mean ± SD across seven lineages (pp), using the same SFT reference throughout. The last row subtracts the $k = 2$ history change from the $k = 3$ change within each lineage.

Threshold	H_B	H_S	H_A	$K_{S/B}$	$K_{A/B}$
$k = 2$	-0.49 ± 3.90	$+3.90 \pm 5.13$	$+3.30 \pm 5.31$	$+4.40 \pm 7.45$	$+3.79 \pm 8.23$
$k = 3$	-0.60 ± 3.57	-2.65 ± 8.95	-14.20 ± 16.89	-2.05 ± 8.06	-13.59 ± 18.48
k3-k2	-0.11 ± 1.76	-6.55 ± 5.98	-17.50 ± 16.63	-6.44 ± 5.03	-17.39 ± 16.40

Let N be the number of other players who contribute. At fixed λ, increasing the Assurance threshold from two to three players means that both partners must contribute, rather than at least one. Under the same distribution of partner actions, this makes project success no more likely, while the success probability needed to justify contributing remains $1/(1 + \lambda)$. Historical evidence about partner participation may therefore matter differently at the two thresholds. Table 24 computes the threshold-by-history differences directly within lineage.

The older-defection test compares 20 defection rounds followed by 20 cooperation rounds (20D+20C) with 40 cooperation rounds (40C). The recent 20 rounds are therefore identical. HL2 hides the older difference: all 376 compared HL2 prompt pairs are identical. HL40 reveals it. Table 25 measures how much the 20D+20C versus 40C difference changes between HL2 and HL40. Both the older-record content and the resulting prompt length can change. The main history-length comparison measures

Table 25: Exploratory older-defection contrast: (20D+20C minus 40C) at HL40 minus the same contrast at HL2, under CONTINUE. Entries are mean ± seven-lineage SD. The contrast measures responses to controlled scripted-history packages.

Game	Base	SFT	Aligned
PGG	-20.79 ± 13.86	-6.86 ± 9.73	-4.14 ± 7.23
CPR	-27.14 ± 15.86	-21.90 ± 18.58	-32.38 ± 27.08
Donation-PD	-9.05 ± 14.04	$+6.67 \pm 12.35$	$+2.50 \pm 13.43$
Volunteer	-12.50 ± 11.02	-3.45 ± 5.16	$+2.02 \pm 10.68$
Assurance	-12.23 ± 15.26	-3.48 ± 11.23	-1.16 ± 2.85
TD	-21.48 ± 6.79	-16.82 ± 27.21	-6.82 ± 25.09
Trust investor	-19.76 ± 14.76	-8.33 ± 14.08	-0.65 ± 1.73
Trust trustee	-5.71 ± 6.76	$+0.30 \pm 1.64$	$+0.48 \pm 1.26$

behavior in trajectories generated by each checkpoint. This controlled comparison changes older records while holding the recent cooperative tail fixed. It reveals game-specific responses to earlier

defection: the strongest post-trained negative response occurs in CPR, while Base responds more negatively in several other games. Omitting CPR leaves the post-trained cross-game effects close to zero. Thus this probe locates the post-trained response to older defections primarily in CPR within the tested histories. These examples were selected from the already completed 1,094-contrast exploration. Its recorded Benjamini–Hochberg false-discovery-rate adjustment (BH) found two contrasts below 0.05; Holm's familywise-error adjustment found none.

K HISTORY VISIBILITY AND FUTURE INSTRUCTIONS IN PGG

The PGG history-visibility experiment in Section 5.2.2 presents SFT and Aligned with the same saved interaction histories, either shown (REAL) or hidden (MASKED), under CONTINUE or FINAL instructions. The four payoff cells are $(n, r) = (3, 1.4), (3, 2.2), (5, 1.8), (5, 4.2)$. Records come from the complete earlier five-seed, all-seat HL40 archives generated separately by SFT and Aligned. Each query generates a new rationale followed by JSON integer contribution 0 or 1.

Design, rates, and weighting. Let $p_s(f, v)$ be the contribution probability at queried stage $s \in \{S, A\}$, future instruction f, and record visibility v. First compare Aligned with SFT: $\delta_{f,v} = 100[p_A(f, v) - p_S(f, v)]$. Then ask how showing records changes that difference: $H_f^{\text{vis}} = \delta_{f,\text{Real}} - \delta_{f,\text{Masked}}$. Finally compare future instructions: $I = H^{\text{vis}}_{\text{Continue}} - H^{\text{vis}}_{\text{Final}}$. All three comparisons are in pp. Within each lineage, we equally weight four payoffs and two source stages. We weight archived rounds 1–20, 21–50, 51–100, and 101–200 by $(0.10, 0.15, 0.25, 0.50)$, proportional to their lengths. Records are sampled randomly across seats within each source/payoff/time group. Appendix P gives the sampling and interval calculations.

MASKED deletes past actions, earned points, and summaries, replacing them with "Previous-round records are not provided." It keeps the current game, role, round number, and future instructions. No such sentence is added to first-round prompts. Hiding these records also shortens the prompt. In Table 6, both stages contribute less on average when records are shown, under both future instructions. The reduction is larger for Aligned under CONTINUE and for SFT under FINAL. The two interval

Table 26: The record-visibility effect by payoff, archived round, and source stage (pp; mean ± SD across seven lineages). H_f^{vis} is the effect of showing records at Aligned minus that effect at SFT; $I = H^{\text{vis}}_{\text{continue}} - H^{\text{vis}}_{\text{final}}$. Each block separates one factor and averages the others with the original relative weights. These are breakdowns of the same experiment.

Payoff / time / source	$H^{\text{vis}}_{\text{continue}}$	$H^{\text{vis}}_{\text{final}}$	I
(3, 1.4)	-4.3 ± 7.9	$+4.9 \pm 18.1$	-9.2 ± 19.1
(3, 2.2)	-6.2 ± 8.9	$+4.9 \pm 16.7$	-11.1 ± 14.7
(5, 1.8)	-5.9 ± 9.0	$+1.8 \pm 16.1$	-7.7 ± 17.0
(5, 4.2)	-6.3 ± 10.8	$+3.5 \pm 17.5$	-9.7 ± 21.3
Rounds 1–20	-6.3 ± 11.0	$+1.8 \pm 12.1$	-8.1 ± 16.3
21–50	-5.0 ± 9.2	$+3.0 \pm 17.2$	-8.0 ± 20.3
51–100	-6.3 ± 9.5	$+3.6 \pm 17.5$	-9.9 ± 18.8
101–200	-5.4 ± 8.2	$+4.5 \pm 17.7$	-9.9 ± 17.1
SFT-source archive	-6.8 ± 9.5	$+3.9 \pm 17.8$	-10.6 ± 18.6
Aligned-source archive	-4.6 ± 8.1	$+3.7 \pm 15.5$	-8.2 ± 16.6

Table 27: Uncertainty in the archived-record comparisons (pp). H_{continue} abbreviates $H^{\text{vis}}_{\text{CONTINUE}}$; ± is SD across seven lineages. The fixed-target interval reflects sampling records and answers from the same models; the bootstrap and t_6 columns use differences among lineages. Each interval is 97.5%, as specified for the two primary comparisons.

Contrast	Mean ± lineage SD	Fixed target	Composition boot.	t_6 sensitivity
H_{continue}	-5.66 ± 8.79	$[-6.62, -4.70]$	$[-11.95, +1.40]$	$[-15.52, +4.21]$
I	-9.42 ± 17.52	$[-10.82, -8.02]$	$[-24.29, +2.45]$	$[-29.08, +10.24]$

types answer different questions. A fixed-archive interval describes the effect of sampling new records and model answers while holding the seven models and archive fixed. A model-composition interval describes variation across the selected lineages. Table 28 also separates two SDs: within each lineage, the SD compares four payoff estimates; the final row compares seven lineage means. The interaction I is negative when averaged separately at each of the four payoffs, each time bin, or each source stage. It is negative in five of seven lineages. The 97.5% model-composition interval is $[-24.29, +2.45]$ pp, so the average comparison remains sensitive to which models are included.

Table 28: Archived-record results for each lineage (pp). δ_{CR} is Aligned's rate minus SFT's under CONTINUE with records shown. H_f is the effect of showing records at Aligned minus that effect at SFT, under frame f; it abbreviates H_f^{vis}. $I = H_{\mathrm{continue}} - H_{\mathrm{final}}$. Each model row gives mean ± SD across four payoff settings; the last row gives mean ± SD across seven lineage means. Source-stage and time-bin weights are as defined in the text.

Lineage	δ_{CR}	H_{continue}	H_{final}	I
DeepSeekMath	$+1.03 \pm 1.35$	-14.87 ± 2.72	-22.56 ± 2.16	$+7.70 \pm 4.10$
InternLM2	$+1.08 \pm 2.48$	-1.83 ± 4.76	-0.05 ± 3.16	-1.78 ± 4.54
Nemotron-Cascade	$+11.50 \pm 1.12$	$+9.00 \pm 1.32$	$+17.19 \pm 8.48$	-8.18 ± 9.21
OLMo-3	-6.34 ± 1.32	-10.16 ± 1.99	$+12.00 \pm 3.43$	-22.16 ± 2.75
SmolLM2	-0.60 ± 1.31	$+0.97 \pm 1.96$	-5.25 ± 3.01	$+6.22 \pm 4.17$
Tülu-3	-1.85 ± 0.89	-14.53 ± 2.16	$+27.58 \pm 6.43$	-42.10 ± 7.27
Zephyr	-8.58 ± 1.59	-8.20 ± 4.57	-2.57 ± 5.33	-5.62 ± 1.61
Equal-lineage mean	-0.54 ± 6.45	-5.66 ± 8.79	$+3.76 \pm 16.55$	-9.42 ± 17.52

Influence of each lineage. Table 29 removes one lineage at a time and recomputes the mean with all other weights unchanged. The mean I stays negative in all seven cases. The FINAL contrast $H_{\mathrm{Final}}^{\mathrm{vis}}$ becomes negative when Tülu-3 is removed, showing that this model affects the ordering under FINAL. The main comparison includes all seven lineages.

Table 29: How the archived-record results change when one lineage is removed. The first row includes all seven lineages; each other row excludes the named lineage. Entries give mean ± SD across the remaining lineages (pp), with the same relative weights for payoffs, source stages, and archived rounds.

Omitted lineage	$H_{\mathrm{Continue}}^{\mathrm{vis}}$	$H_{\mathrm{Final}}^{\mathrm{vis}}$	I
None (primary)	-5.66 ± 8.79	$+3.76 \pm 16.55$	-9.42 ± 17.52
DeepSeekMath	-4.12 ± 8.54	$+8.15 \pm 12.93$	-12.27 ± 17.32
InternLM2	-6.30 ± 9.45	$+4.40 \pm 18.04$	-10.69 ± 18.84
Nemotron-Cascade	-8.10 ± 6.53	$+1.52 \pm 16.93$	-9.63 ± 19.19
OLMo-3	-4.91 ± 9.38	$+2.39 \pm 17.69$	-7.30 ± 18.18
SmolLM2	-6.76 ± 9.08	$+5.26 \pm 17.60$	-12.03 ± 17.65
Tülu-3	-4.18 ± 8.62	-0.21 ± 14.02	-3.97 ± 10.92
Zephyr	-5.23 ± 9.55	$+4.82 \pm 17.87$	-10.05 ± 19.11

L TEMPORAL-FRAME EFFECTS BY STAGE AND VISIBLE-HISTORY LENGTH

Table 5 reports the overall absolute cooperation rates under FINAL and CONTINUE. Table 30 separates their difference by training stage and visible-history length.

Table 30: The effect of CONTINUE rather than FINAL, shown separately for each stage and history length (pp; mean ± SD across seven lineages). At each nonzero history length, we average payoff settings and seven scripted behavior patterns equally. HL0 shows no past records. Each prompt asks for one current action.

Game	Stage	HL0	HL2	HL10	HL40
PGG	Base (B)	$+1.4 \pm 14.4$	-0.8 ± 5.2	-1.5 ± 5.4	$+1.8 \pm 4.3$
	SFT (S)	-2.8 ± 11.8	-2.6 ± 9.9	$+6.4 \pm 8.6$	$+4.6 \pm 5.5$
	Aligned (A)	$+1.9 \pm 8.3$	-7.2 ± 7.0	-1.4 ± 9.5	-3.8 ± 7.9
CPR	Base (B)	$+4.3 \pm 13.9$	$+9.2 \pm 6.7$	$+2.1 \pm 2.7$	$+7.6 \pm 5.1$
	SFT (S)	$+9.4 \pm 22.2$	$+16.3 \pm 9.5$	$+15.5 \pm 9.9$	$+10.9 \pm 14.5$
	Aligned (A)	$+9.0 \pm 13.5$	$+25.2 \pm 16.6$	$+26.0 \pm 24.7$	$+26.5 \pm 32.8$
Donation-PD	Base (B)	$+7.3 \pm 8.8$	-3.7 ± 11.6	-1.2 ± 8.6	$+1.5 \pm 8.6$
	SFT (S)	$+6.7 \pm 16.8$	$+9.9 \pm 20.5$	$+3.6 \pm 15.1$	$+2.7 \pm 8.5$
	Aligned (A)	-5.2 ± 8.8	$+3.9 \pm 18.7$	$+3.4 \pm 13.8$	$+7.2 \pm 11.2$
Volunteer	Base (B)	-4.8 ± 11.5	$+3.7 \pm 6.7$	$+3.1 \pm 5.5$	$+5.0 \pm 7.6$
	SFT (S)	-0.4 ± 9.9	$+2.3 \pm 12.5$	$+5.9 \pm 13.2$	$+0.6 \pm 10.9$
	Aligned (A)	$+2.0 \pm 5.2$	$+0.2 \pm 7.1$	$+7.0 \pm 17.0$	$+1.3 \pm 12.9$
Assurance	Base (B)	$+6.2 \pm 13.4$	$+0.5 \pm 8.9$	-0.7 ± 8.1	$+2.6 \pm 10.9$
	SFT (S)	-8.8 ± 17.6	-4.5 ± 17.6	$+4.3 \pm 19.1$	$+13.3 \pm 18.6$
	Aligned (A)	-6.2 ± 12.0	-5.1 ± 11.2	$+3.3 \pm 17.5$	$+7.9 \pm 16.0$
TD	Base (B)	-15.7 ± 16.0	-7.7 ± 8.4	-8.4 ± 10.6	-7.9 ± 9.5
	SFT (S)	-19.4 ± 23.3	-9.6 ± 7.2	-10.6 ± 8.4	-13.0 ± 6.8
	Aligned (A)	-24.8 ± 18.2	-12.0 ± 14.5	-20.3 ± 14.8	-20.4 ± 15.7
Trust	Base (B)	$+3.6 \pm 12.0$	-4.0 ± 7.3	$+0.4 \pm 4.2$	$+2.4 \pm 4.3$
	SFT (S)	$+3.2 \pm 10.9$	-2.1 ± 16.2	$+5.1 \pm 6.0$	$+6.7 \pm 9.4$
	Aligned (A)	$+1.4 \pm 3.0$	-0.1 ± 13.2	$+1.8 \pm 10.7$	$+2.8 \pm 5.5$

M PROMPTS, INFERENCE SETTINGS, AND DATA SOURCES

Table 31 summarizes which models, payoffs, and decisions each experiment uses. Because SmolLM2's SFT release names an 8k-context parent rather than the evaluated Base release, we also calculate the history contrasts without this lineage: $K_{S/B} = -5.40 \pm 6.63$ and $K_{A/B} = -8.37 \pm 7.12$ pp (six-lineage SD). The main results keep all seven lineages. Parts of the history-length experiment code are reused from Liu et al. (2026); all data, figures, results, and tables in this paper are new.

Table 31: Study scopes. B/S/A denotes Base/SFT/Aligned. Counts identify the stated sampling or configuration level.

Study	Stages, design and interface	Sampling and temporal condition
Canonical 84	B/S/A; 84 (21 checkpoints); All seats within each trajectory; symbolic, rationale + action; homogeneous endogenous self-play; OS: 1; repeated: 200, HL2/40	OS no future text; repeated 0.99 / long-run / visible later; 20 OS IDs; 5 repeated seeds; original + new CPR (post-hoc amendment); main analysis
Original 82	B/S/A; 82 (21 checkpoints); all seats; symbolic, rationale + action; original homogeneous self-play; OS: 1; repeated: 200, HL2/40	same canonical frame definitions; 20 OS IDs; 5 repeated seeds; original absolute-surface reference; sensitivity analysis
PGG source histories	S/A; 4 payoffs; all-seat HL40 records from homogeneous self-play; numeric JSON contribution; rationale + action; 200 recorded rounds	0.99 / long-run; 5 trajectory seeds; earlier archive supplies fixed inputs for the history-visibility experiment
Continuation	B/S/A; 82 / 94 role settings; CPR 12; focal Player 1; Trust investor / representative trustee; action-only game-native JSON; seven scripted patterns, plus no-history; current round 1 or 41; HL0/2/10/40	CONTINUE vs FINAL; contradiction secondary; 20 IDs per prompt; cross-game future-instruction comparison; Base secondary interface diagnostic

Study	Stages, design and interface	Sampling and temporal condition
PGG history visibility	S/A; 4; two source stages; archival seats selected by simple random sampling without replacement; numeric JSON; rationale + action; old complete HL40 archives; rounds 1–200, four time bins	CONTINUE/FINAL × REAL/MASKED; 8 records/stratum; 32 draws/query; completed follow-up runtime; fixed-archive explanatory
Nemotron case	Seven checkpoints of one lineage; common original82; all seats; rationale + action; OS and 200-round HL2/40	20 OS draws; five repeated seeds; Aligned is RLHF; final case stage is SWE-RL; Appendix T
Controlled decision-value case study	All 21 checkpoints; PGG, Donation-PD and Assurance; 16 configurations with neutral prompts; focal model against two programmed partners with disclosed rules	Fixed finite individual-return objective; current actions scored with optimal continuation; original neutral prompts; Appendix U

Interaction and decision protocol. At every canonical round, all fixed seats use the same checkpoint. The runner first assembles seat-specific prompts from the same completed joint actions and own payoff records, collects all seats concurrently, and only then appends the new joint round. At round t, canonical prompts list the most recent $\min(L, t-1)$ joint actions and the focal player's payoff in each listed round. They retain the current round number, fixed seat, and game rules, but include neither a cumulative-payoff total nor a model-written history summary. Canonical OS samples 20 one-round, all-seat draws. Neither canonical repeated prompts nor the PGG archive-generation prompts announce the logging cap of 200 rounds: they state 99% continuation, future observability and a long-run points objective. The canonical runner records a fixed 200-round window without sampling a stopping time from that stated probability. FINAL explicitly states that no later round will occur.

Canonical rationale requests contain one user message, followed for action selection by that prompt, the queried checkpoint's own rationale as an assistant message, and a user action instruction. No explicit system message is inserted by this client; tokenizer-template defaults remain a separate serving property. Sampling is temperature 0.7, top-p 0.8, top-k 20; rationale budget is 256 tokens and the symbolic action budget is eight. The PGG history-visibility experiment and its source histories instead use strict JSON with integer contribution 0 or 1 and a 32-token action budget. The continuation control uses one action-only, game-native JSON request capped at 32 tokens.

Canonical rationale, action, and retry draws use distinct deterministic seeds. The history-visibility experiment assigns unique seeds to each query, stage, and attempt; static draws and trajectory seeds have separate sampling identities. New CPR deployment uses bfloat16 without declared quantization, a 4096-token model limit, no prefix caching, and eager execution. OLMo/Tülu use vLLM 0.12.0; the other new CPR lineages use 0.8.5 with xgrammar. No explicit stop or ignore-EOS override appears in the client. Historical transformers/tokenizers package versions and actual tokenized prompts are not recoverable for every run.

Appendix N explains how we checked the new CPR prompts and handled incomplete runs.

Settings for the shared-record tests. The PGG shown-versus-hidden test uses four payoff settings, $(n, r) = (3, 1.4), (3, 2.2), (5, 1.8), (5, 4.2)$, from its original design. These differ from the main PGG grid. Here "held-out" means previously unobserved payoff cases; the initial two lineages were not chosen without knowing earlier outcomes.

The separate final-round/continuing test uses 12 CPR configurations, compared with 16 in the main evaluation. For Trust, it weights the investor and one representative trustee equally; the main evaluation instead averages all three players. It also weights all 22 scripted-history conditions equally, so HL0, HL2, HL10, and HL40 receive total weights $1/22, 7/22, 7/22, 7/22$.

The seven scripted-history patterns. Each pattern specifies 40 completed rounds: (1) everyone cooperates throughout; (2) everyone takes the noncooperative action throughout; (3) the focal player cooperates while all partners do not; (4) the focal player does not cooperate while all partners do; (5) players alternate actions each round, with odd- and even-numbered seats taking opposite actions; (6) everyone cooperates for 20 rounds, then only the focal player continues cooperating; (7) everyone takes the noncooperative action for 20 rounds, then everyone cooperates. Binary games use

their designated cooperative action and its alternative; TD uses the highest and lowest legal claims, respectively. Each prompt shows the last 2, 10, or 40 rounds of that same record, including the actions and focal player's payoffs. HL0 shows no records. All conditions ask for a decision in round 41.

Before this test collected any decisions, a startup failure required reducing the serving context limit to 4096 tokens. For SmolLM2 Base's final 59 missing decisions, the retry limit was raised from four to 32 attempts. Base results in this test help assess the response interface; the primary stage comparison is Aligned versus SFT.

Raw-log coverage. Raw-log coverage is more complete for the new CPR and follow-up experiments than for the original runs.

M.1 MAIN-EXPERIMENT PROMPT TEMPLATE

The template below reproduces the main experiment's PGG wording for Player 1 with $n = 3$, multiplier $r = 1.8$, and $\mathrm{MPCR} = 0.6$. Square brackets mark dynamically filled fields; line wrapping is adapted for display. The same two-request structure is used across released stages. Other games replace the rules and legal actions with their game- and role-specific versions (Appendix C).

First request: rationale. The client sends the following as a single user message.

```
You are Player 1 in round [round] of a repeated game with
the same 2 other player(s).

There are 3 players, each with an endowment of 1.  Let k be
the number choosing A0 (contribute the endowment).  The
total contribution is multiplied by 1.8 and split equally,
so every player receives 1.8*k/3 from the public good.  An
A1 player also keeps their 1-point endowment; an A0 player
does not.  Thus contributing raises total welfare but lowers
the contributor's own payoff holding all other actions
fixed.

Your action will be visible to the same other player(s) in
future rounds and may influence their decisions.
After each round there is a 99% chance of another round.
Maximize your total points in the long run.
You can see at most the [L] most recent completed round(s).

Past rounds:
[history]

Available actions:  A0 or A1.
Briefly state your strategic rationale in at most 120 words.
This is the rationale stage; a separate constrained
action-selection step will follow, so do not add a special
final-answer line.
```

History field and one-shot variant. Here `[round]` is the current round number and `[L]` is 2 or 40. The `[history]` field lists the most recent $\min(L, t-1)$ completed rounds in chronological order, one line per round, using this format for Player 1:

```
Round [past round]:  You=[own action], Player 2=[action],
Player 3=[action]; your payoff=[payoff].
```

At the first repeated round, this field instead reads `(This is the first round; no past history is available.)`; the future-interaction instructions remain. For one-shot decisions, the opening becomes `You are Player 1 in a decision-making game.` and the entire block from `Your action will be visible` through the history field is omitted. Rules, available actions, and the rationale instruction are retained.

Second request: constrained action. The next request contains the original user message, that checkpoint's generated rationale as an assistant message, and the following new user message:

```
Now commit to one action for this round.  Return only one
of:  A0, or A1.  Do not include any explanation.
```

Decoding constrains the action to `A0` or `A1`. The rationale request's 120-word instruction is separate from its 256-token generation cap; action generation is capped at eight tokens. The client inserts no explicit system message. The shared-record tests use their separately specified JSON interfaces rather than this action-label template.

N FAILED REQUESTS, RETRIES, AND BOUNDS ON MISSING ACTIONS

For the new CPR runs, we reconstructed all 1,018,080 action prompts from the completed rounds that preceded them. Logged prompt lengths range from 133 to 1,994 tokens, and prompts contain the full two- or forty-round history when enough rounds are available. This check compares prompt text and recorded lengths; it does not reconstruct unavailable server-tokenization or historical truncation logs.

New 0.4/0.6 runs use the same game factory and original five trajectory seeds or 20 one-shot draw IDs at all 21 checkpoints and all three exposures. In total these eight new configurations require 1,680 repeated trajectories and 3,360 one-shot draws, or 1,018,080 logical action decisions. These parameter points form a post-hoc grid amendment in a separately recorded serving cohort. Original 0.5 observations remain in the 82-grid archive regardless of their results.

All 5,040 new CPR replicate rows are complete, including 80 preserved-prefix repairs. Zephyr Base and DeepSeekMath Base allow 64 attempts independently per rationale and action step. One-shot previously allowed 12 attempts per step; repeated play allowed three. Rejection of empty rationales can depend on the sampled model output. The enlarged cap therefore defines a distinct serving procedure, assessed below through conservative suffix bounds and recorded cohort provenance.

Retry prompts contain only completed past rounds, not current or future actions. We keep valid decisions already collected within the current round and record failed attempts. We use the first valid response: a nonempty rationale followed by a legal action. Retries advance the seed by a fixed rule and do not select among successful responses according to the chosen action. The prompt, rationale, temperature, and output format are unchanged within a retry sequence. Logged failures distinguish empty rationales, request failures, and parsing failures. The client does not separately record automatic HTTP retries inside the SDK.

Actual beyond-old-cap use occurs in 153 repeated Base trajectories: Zephyr HL2/HL40 $40/40$ each; DeepSeekMath HL2 $33/40$ and HL40 $40/40$. No new one-shot draw exceeds the old 12-step cap. There are 3,725 attempts beyond old per-step caps; the first affected round ranges from 1 to 186 (median 18). These beyond-cap events concern empty-rationale sampling.

To measure sensitivity to the larger retry limit, keep only rounds before the first request that exceeded the old limit, and allow every later action to be either cooperative or noncooperative. Let c be the observed cooperative actions, m all remaining actions and N the full trajectory action count. Any legal completion lies in $[c/N, (c+m)/N]$. Treating the entire first affected round as unknown is conservative for seats already completed. For a linear estimand $a^T p$, let $p_j \in [l_j, u_j]$ be the cell outcome and a_j its fixed contrast weight. The lower bound sums $a_j l_j$ for $a_j \geq 0$ and $a_j u_j$ otherwise; the upper bound reverses those endpoints. For percentage outcomes, OLS drop weights are $a_j = -(z_j - \bar{z}) / \sum_k (z_k - \bar{z})^2$, where $\bar{z}$ is the mean of the tested normalized coordinates, including the nonuniform original and union grids. Nuisance, game and lineage weights remain unchanged. These worst-case completion bounds address the unknown suffix after a beyond-cap event. Their scope is completion sensitivity; runtime-cohort differences from round 1 require a separate comparison. All completed trajectories remain in the stated panel.

Table 32: How much could results change after a request exceeded the original retry limit? Bounds allow every subsequent action to take either legal value, with all comparison weights unchanged (pp). They assess the effect of these unknown actions; the text separately describes deployment differences.

Primary summary	Observed mean	Completion range
$K_{S/B}$	-6.93	[-8.53, -4.83]
$K_{A/B}$	-9.76	[-11.36, -7.66]
$S_S - S_B$ HL40	+2.83	[+1.58, +4.04]
$Q_{S/B}$	+1.07	[-1.16, +3.28]
$Q_{A/B}$	+2.34	[+0.11, +4.55]

Table 33: History changes for each lineage, averaged over the six binary games. Each entry is mean ± SD across five matched run IDs, after averaging payoff configurations within game and games equally. This SD describes repeated runs of one lineage, rather than differences among models.

Lineage	H_B	H_S	H_A	$K_{S/B}$	$K_{A/B}$	$K_{A/S}$
OLMo	$+1.7 \pm 1.4$	-13.5 ± 1.4	-8.9 ± 0.8	-15.2 ± 2.8	-10.7 ± 2.0	$+4.6 \pm 1.0$
SmolLM2	$+0.0 \pm 0.1$	-16.1 ± 2.4	-18.1 ± 2.3	-16.2 ± 2.4	-18.1 ± 2.4	-1.9 ± 2.0
DeepSeekMath	$+0.9 \pm 0.2$	-5.6 ± 0.8	-9.7 ± 1.1	-6.5 ± 0.8	-10.7 ± 1.0	-4.1 ± 1.2
InternLM2	-7.6 ± 0.8	-2.3 ± 0.4	-2.8 ± 0.5	$+5.3 \pm 0.9$	$+4.8 \pm 0.6$	-0.5 ± 0.6
Nemotron	-6.0 ± 1.7	-10.9 ± 1.5	-19.7 ± 2.7	-4.9 ± 2.9	-13.7 ± 3.8	-8.8 ± 3.0
Tulu	$+1.4 \pm 0.8$	-6.0 ± 0.9	-12.9 ± 1.1	-7.4 ± 0.9	-14.3 ± 0.3	-6.9 ± 1.0
Zephyr	$+1.3 \pm 0.6$	-2.4 ± 0.2	-4.4 ± 0.8	-3.7 ± 0.6	-5.7 ± 1.0	-2.0 ± 0.7

Table 34: Payoff-response differences for each lineage (pp). Entries give mean ± SD across 20 one-shot draws or five repeated runs, pairing the same nominal IDs for differences. Each run first averages the other payoff parameters, then the six symmetric games equally, including TD and excluding Trust. Q measures how the response difference changes from HL2 to HL40.

Lineage	$S_S - S_B$ OS	HL2	HL40	$Q_{S/B}$	$Q_{A/B}$
OLMo	$+0.3 \pm 7.0$	-0.7 ± 1.1	-7.5 ± 7.5	-6.8 ± 7.5	-6.7 ± 4.3
SmolLM2	$+0.6 \pm 5.7$	$+0.9 \pm 1.2$	$+0.6 \pm 2.2$	-0.3 ± 3.3	-0.4 ± 3.1
DeepSeekMath	-0.0 ± 8.1	$+3.0 \pm 1.1$	$+3.6 \pm 2.1$	$+0.5 \pm 2.4$	$+3.3 \pm 1.1$
InternLM2	$+2.7 \pm 11.7$	$+3.3 \pm 1.0$	$+6.8 \pm 1.4$	$+3.5 \pm 2.3$	$+3.7 \pm 1.9$
Nemotron	$+3.9 \pm 8.0$	-0.2 ± 1.4	$+5.5 \pm 5.8$	$+5.6 \pm 5.6$	$+6.7 \pm 7.4$
Tulu	$+1.9 \pm 5.4$	$+2.8 \pm 1.1$	$+7.0 \pm 2.4$	$+4.2 \pm 3.2$	$+7.4 \pm 4.5$
Zephyr	$+1.0 \pm 11.5$	$+3.2 \pm 0.8$	$+3.9 \pm 0.9$	$+0.7 \pm 1.2$	$+2.4 \pm 1.2$

O Results for Each Model Lineage

Tables 33 and 34 report SDs across runs of each checkpoint. Differences match runs by replicate ID and use the same game and payoff weights throughout. These SDs describe variation across repeated runs of one model; the main tables' cross-lineage SDs instead describe differences among seven models.

P Sampling Archived Records and Estimating Uncertainty

The PGG history-visibility experiment (Section 5.2.2; Appendix K) has two sources of sampling variation: which saved records are selected and which answers the models generate for them. We independently sample eight records without replacement from each lineage × payoff × source stage × time-bin stratum of the complete old HL40 archive. The sampling population includes coordinates used in earlier experiments. For a bin with B rounds and n_p seats, the original five-seed archive has $N = 5Bn_p$ records; bins 1–20, 21–50, 51–100 and 101–200 have weights 0.10, 0.15, 0.25 and 0.50. Within a lineage/payoff, source stages receive weight $1/2$; payoffs and lineages receive $1/4$ and $1/7$. Each record is queried under eight stage/frame/visibility arms, with 32 independent full rationale/action draws per query. The fixed-panel total is 14,336 queries and 458,752 decisions.

For sampled record i, let $\hat{\boldsymbol{p}}_i \in \mathbb{R}^8$ contain the estimated cooperation probabilities for the two queried stages crossed with the two future instructions and two history-visibility conditions. Each entry uses $d = 32$ independent complete rationale-and-action draws. Let $\hat{V}_i$ be the diagonal estimated covariance matrix of these eight query means; an entry with estimated probability $\hat{p}$ has variance estimate $\hat{p}(1-\hat{p})/(d-1)$. A stratum contains N archived records and contributes $n_{\text{rec}} = 8$ records sampled without replacement. Define its sampling fraction $f = n_{\text{rec}}/N$, the sample mean $\bar{\hat{\boldsymbol{p}}} = n_{\text{rec}}^{-1} \sum_i \hat{\boldsymbol{p}}_i$, and the sample covariance $S_{\hat{\boldsymbol{p}}}$ of the record-level vectors, using denominator $n_{\text{rec}} - 1$. The reused covariance estimator is

$$\widehat{\text{Cov}}(\bar{\hat{\boldsymbol{p}}}) = \frac{1-f}{n_{\text{rec}}} S_{\hat{\boldsymbol{p}}} + \frac{f}{n_{\text{rec}}^2} \sum_{i=1}^{n_{\text{rec}}} \hat{V}_i.$$

Independent stratum contributions are combined using squared aggregation weights, giving the aggregate covariance matrix $\hat{\Sigma}$. For a scalar comparison specified by coefficient vector $\boldsymbol{c}$, its estimated variance is $\boldsymbol{c}^\top \hat{\Sigma} \boldsymbol{c}$. The record-level covariance preserves dependence induced by the same sampled input across query conditions. Normal-approximation intervals and calibration simulations use this covariance estimator. For the two primary contrasts, $H^{\mathrm{vis}}_{\mathrm{continue}}$ and I, we report componentwise 97.5% normal intervals. $H^{\mathrm{vis}}_{\mathrm{final}}$ uses the supportive 95% interval. Sampling SE and model-composition SD have different targets.

Calibration uses six synthetic probability settings: interior cooperation probabilities with low between-seed variation, high between-seed variation, or high between-record variation; probabilities near zero; probabilities near one; and a rare high-cooperation seed mode with probability 0.10. We test each setting with one and seven lineages, generating 2,000 repeated samples per combination (24,000 synthetic datasets). Within each combination, a five-seed archive and its cooperation probabilities remain fixed. Every sample selects eight records per stratum without replacement and generates 32 independent Bernoulli actions per query condition, retaining the payoff, source-stage, and time-bin structure and weights described above. We apply the same normal-interval estimator to the two primary contrasts and compare with their exact fixed-archive means. Coverage ranges 0.9405–0.9545 at nominal 0.95 and 0.9655–0.9785 at nominal 0.975, within three Monte Carlo standard errors of the nominal levels across these scenarios.

Q Complete Canonical Configuration Tables

What is a configuration? A configuration fixes a game's player count and every payoff parameter. It does not specify a model, training stage, or history length: the same configuration is evaluated at all three stages under one-shot, HL2, and HL40 conditions. Table 35 therefore has three rows per configuration. The 84 configurations comprise 8 PGG, 16 CPR, 12 Donation-PD, 12 Volunteer, 8 Assurance, 16 TD, and 12 Trust settings. Appendix C gives their payoff matrices and functions.

How to read the configuration IDs. The prefix identifies the game. In `pgg3`, `pgg5`, `cpr3`, `pd3`, `vd3`, `sh3`, and `tg3`, the number is the player count. In `td8`, `td16`, `td32`, and `td64`, it instead gives the number of legal claim levels; TD always has two players. The letter `p` inside a number represents a decimal point. Other decimal encodings are game-specific, as explained below. The suffix `neutral` identifies the descriptive payoff-based wording of the game prompt, not an additional payoff value or training stage.

- **PGG.** In `pgg3_r12`, there are $n = 3$ players and the contribution pool is multiplied by $r = 1.2$, giving MPCR $\alpha = r/n = 0.4$. The digits after `r` encode ten times the pool multiplier: for example, `r24` means 2.4. The grid crosses $n \in \{3, 5\}$ with $\alpha \in \{0.4, 0.5, 0.6, 0.8\}$, yielding eight configurations.
- **CPR.** In `cpr3_h1p25_t20`, high extraction is $h = 1.25$ and $\theta = 0.20$; restrained extraction is always 1. The `t20`, `t40`, `t60`, and `t80` codes mean $\theta = 0.2, 0.4, 0.6, 0.8$. Crossing these with $h \in \{1.25, 1.5, 2.5, 3\}$ gives 16 configurations. Together, h and θ set the resource-return parameter $V = 3h + 1 + 2\theta$.
- **Donation-PD.** In `pd3_c02_r11_neutral`, donating costs $c = 0.2$ and the total-benefit-to-cost ratio is $R = 1.1$, so each other player receives $b = Rc/2 = 0.11$. Here `r` encodes R, not the PGG pool multiplier. The cost codes `c02`, `c06`, and `c12` mean $c = 0.2, 0.6, 1.2$; `r11`, `r14`, `r20`, and `r24` mean $R = 1.1, 1.4, 2.0, 2.4$. Their combinations give 12 configurations.
- **Volunteer.** In `vd3_b4_c15`, provision gives every player benefit $B = 4$ and each volunteer pays $c = 1.5$, giving relative cost $\rho = c/B = 0.375$. The grid crosses $B \in \{2, 4, 8\}$ with $\rho \in \{0.125, 0.375, 0.625, 0.875\}$. Corresponding costs are $\{0.25, 0.75, 1.25, 1.75\}$ for $B = 2$, $\{0.5, 1.5, 2.5, 3.5\}$ for $B = 4$, and $\{1, 3, 5, 7\}$ for $B = 8$. For example, `c025`, `c1p25`, and `c1` mean 0.25, 1.25, and 1, respectively. These are 12 configurations.
- **Assurance.** The prefix `sh3` denotes the three-player Assurance game. In `sh3_k2_h3p15`, at least $k = 2$ contributors are required and each successful contributor receives $h = 3.15$. The sure outside payoff is 3, so the successful-project premium is $\lambda =$

$(h-3)/3 = 0.05$. Crossing $k \in \{2,3\}$ with $h \in \{3.15, 3.3, 3.6, 4.2\}$, equivalently $\lambda \in \{0.05, 0.10, 0.20, 0.40\}$, gives eight configurations.

- **TD.** In `td16_tau025_neutral`, there are $K = 16$ legal claims, from 2 through 17, and $\tau = 0.025$. The bonus/penalty is $1 + \tau(K-1) = 1.375$. The codes `tau025`, `tau05`, `tau10`, and `tau20` mean $\tau = 0.025, 0.05, 0.10, 0.20$. Crossing these with $K \in \{8, 16, 32, 64\}$ gives 16 configurations.
- **Trust.** In `tg3_m2_s02_neutral`, the investment multiplier is $m = 2$ and the return parameter is $s = 0.2$. Investing the 10-point endowment gives each trustee $5m = 10$ points; a returning trustee transfers $5(1+s) = 6$ points to the investor. The `s02`, `s04`, `s06`, and `s08` codes mean $s = 0.2, 0.4, 0.6, 0.8$. Crossing these with $m \in \{2, 3, 4\}$ gives 12 configurations.

How to read the results. Within each configuration, *one-shot* has neither history nor future-round statements; HL2 and HL40 are separate 200-round repeated-game conditions showing up to two or forty past rounds. Columns compare Base, SFT, and Aligned (A). Each entry is the mean ± SD across seven lineage-level rates, after averaging the 20 one-shot draws or five repeated-game seeds within each lineage. Binary-game entries are cooperation percentages; TD entries are normalized claims multiplied by 100. All players contribute to the reported rate, including the investor and two trustees in Trust.

Table 35: All 84 main configurations, all three conditions and stages. The preceding guide decodes each game's configuration IDs; Appendix C gives the full payoff rules. Percent rates/TD normalized scores are mean ± SD across seven lineages. A denotes Aligned.

Configuration	Exposure	Base	SFT	A
pgg3_r12	one-shot	78.57 ± 16.48	49.05 ± 13.54	57.38 ± 15.00
pgg3_r12	HL2	61.99 ± 37.25	40.67 ± 15.52	45.04 ± 24.55
pgg3_r12	HL40	56.83 ± 41.79	29.89 ± 16.06	30.40 ± 29.10
pgg3_r15	one-shot	77.38 ± 15.66	49.29 ± 11.97	60.95 ± 13.26
pgg3_r15	HL2	61.86 ± 37.49	42.72 ± 17.62	46.65 ± 25.10
pgg3_r15	HL40	56.50 ± 42.37	33.29 ± 16.49	36.08 ± 30.05
pgg3_r18	one-shot	77.14 ± 18.82	46.67 ± 14.34	56.90 ± 14.95
pgg3_r18	HL2	62.67 ± 36.97	43.19 ± 17.26	46.30 ± 24.82
pgg3_r18	HL40	58.41 ± 40.09	37.30 ± 17.56	37.00 ± 29.28
pgg3_r24	one-shot	80.00 ± 16.44	55.95 ± 12.35	65.00 ± 10.80
pgg3_r24	HL2	63.97 ± 35.07	44.85 ± 16.10	48.67 ± 24.11
pgg3_r24	HL40	60.25 ± 37.24	42.42 ± 19.55	39.56 ± 30.41
pgg5_r20	one-shot	78.71 ± 19.01	51.57 ± 10.21	58.43 ± 9.88
pgg5_r20	HL2	60.69 ± 37.16	44.45 ± 13.95	48.42 ± 23.10
pgg5_r20	HL40	57.60 ± 41.77	37.51 ± 21.89	32.45 ± 29.25
pgg5_r25	one-shot	80.43 ± 15.20	49.57 ± 10.33	58.29 ± 11.73
pgg5_r25	HL2	59.88 ± 38.34	44.13 ± 13.78	48.34 ± 22.42
pgg5_r25	HL40	57.61 ± 41.43	37.57 ± 16.32	35.03 ± 27.91
pgg5_r30	one-shot	77.43 ± 19.97	51.00 ± 9.49	59.29 ± 11.76
pgg5_r30	HL2	60.91 ± 36.83	46.83 ± 15.23	51.07 ± 23.59
pgg5_r30	HL40	58.71 ± 40.44	39.83 ± 19.84	40.40 ± 33.20
pgg5_r40	one-shot	77.57 ± 19.40	54.14 ± 7.99	60.00 ± 12.01
pgg5_r40	HL2	60.71 ± 37.08	47.99 ± 15.70	51.38 ± 22.74
pgg5_r40	HL40	59.39 ± 39.01	44.01 ± 22.75	44.77 ± 33.71
cpr3_h1p25_t20	one-shot	71.43 ± 22.58	33.33 ± 8.98	35.24 ± 19.16
cpr3_h1p25_t20	HL2	68.17 ± 25.87	38.40 ± 16.85	34.40 ± 15.17
cpr3_h1p25_t20	HL40	65.36 ± 32.39	30.25 ± 17.17	22.40 ± 16.23
cpr3_h1p25_t40	one-shot	71.67 ± 18.71	41.90 ± 19.71	44.76 ± 12.56
cpr3_h1p25_t40	HL2	63.74 ± 21.21	39.90 ± 20.61	39.85 ± 13.77
cpr3_h1p25_t40	HL40	62.45 ± 25.11	31.62 ± 19.70	23.03 ± 12.92
cpr3_h1p25_t60	one-shot	69.76 ± 19.59	41.19 ± 18.02	47.38 ± 13.15
cpr3_h1p25_t60	HL2	62.85 ± 21.87	39.21 ± 18.90	38.04 ± 11.63
cpr3_h1p25_t60	HL40	61.05 ± 27.31	28.95 ± 19.41	21.82 ± 12.42
cpr3_h1p25_t80	one-shot	69.52 ± 21.81	34.76 ± 9.40	37.62 ± 18.05
cpr3_h1p25_t80	HL2	67.05 ± 26.65	38.97 ± 19.63	33.59 ± 15.50
cpr3_h1p25_t80	HL40	64.90 ± 32.25	25.14 ± 15.28	16.51 ± 12.63

Configuration	Exposure	Base	SFT	A
cpr3_h1p5_t20	one-shot	69.76 ± 20.51	33.81 ± 8.03	40.00 ± 24.00
cpr3_h1p5_t20	HL2	66.94 ± 27.94	40.68 ± 19.00	38.06 ± 15.96
cpr3_h1p5_t20	HL40	69.34 ± 27.98	31.49 ± 19.17	21.43 ± 14.75
cpr3_h1p5_t40	one-shot	70.48 ± 22.08	39.76 ± 19.13	39.29 ± 11.90
cpr3_h1p5_t40	HL2	63.11 ± 23.25	41.70 ± 19.64	41.91 ± 12.56
cpr3_h1p5_t40	HL40	64.57 ± 24.71	33.38 ± 21.42	27.95 ± 15.70
cpr3_h1p5_t60	one-shot	68.57 ± 24.62	40.95 ± 17.66	40.71 ± 14.65
cpr3_h1p5_t60	HL2	62.20 ± 23.83	41.72 ± 19.53	41.55 ± 12.50
cpr3_h1p5_t60	HL40	61.86 ± 28.21	29.70 ± 19.30	22.66 ± 13.72
cpr3_h1p5_t80	one-shot	70.95 ± 19.86	35.71 ± 10.09	37.62 ± 21.66
cpr3_h1p5_t80	HL2	66.50 ± 27.91	38.63 ± 16.33	36.57 ± 14.41
cpr3_h1p5_t80	HL40	62.34 ± 35.35	27.20 ± 15.81	18.11 ± 12.93
cpr3_h2p5_t20	one-shot	68.33 ± 24.98	36.90 ± 13.86	40.24 ± 22.35
cpr3_h2p5_t20	HL2	67.71 ± 25.80	46.99 ± 16.33	41.45 ± 11.72
cpr3_h2p5_t20	HL40	66.20 ± 30.80	31.05 ± 13.89	25.23 ± 13.36
cpr3_h2p5_t40	one-shot	66.90 ± 24.90	39.29 ± 17.90	44.05 ± 13.50
cpr3_h2p5_t40	HL2	62.48 ± 22.55	47.89 ± 16.51	45.85 ± 6.07
cpr3_h2p5_t40	HL40	62.33 ± 24.35	34.38 ± 15.94	27.35 ± 11.32
cpr3_h2p5_t60	one-shot	70.24 ± 24.82	38.57 ± 15.32	44.76 ± 15.26
cpr3_h2p5_t60	HL2	62.02 ± 23.11	47.00 ± 16.67	46.00 ± 6.40
cpr3_h2p5_t60	HL40	61.80 ± 25.54	33.27 ± 14.28	27.38 ± 9.95
cpr3_h2p5_t80	one-shot	69.52 ± 22.56	35.48 ± 15.63	41.19 ± 24.72
cpr3_h2p5_t80	HL2	66.63 ± 26.29	44.53 ± 13.74	41.53 ± 12.26
cpr3_h2p5_t80	HL40	70.07 ± 25.46	26.86 ± 12.54	22.23 ± 11.92
cpr3_h3_t20	one-shot	69.29 ± 24.61	31.43 ± 10.60	40.95 ± 25.87
cpr3_h3_t20	HL2	69.59 ± 24.17	45.63 ± 15.33	42.02 ± 13.11
cpr3_h3_t20	HL40	71.07 ± 27.53	32.76 ± 17.92	25.44 ± 13.40
cpr3_h3_t40	one-shot	68.10 ± 21.76	38.33 ± 18.81	41.19 ± 20.92
cpr3_h3_t40	HL2	63.92 ± 21.25	44.75 ± 16.44	44.61 ± 6.13
cpr3_h3_t40	HL40	65.46 ± 21.11	33.27 ± 15.27	26.42 ± 10.84
cpr3_h3_t60	one-shot	67.38 ± 24.62	37.86 ± 20.77	38.81 ± 17.50
cpr3_h3_t60	HL2	63.25 ± 21.86	44.42 ± 14.68	43.89 ± 7.02
cpr3_h3_t60	HL40	64.48 ± 23.73	30.27 ± 14.61	26.50 ± 9.30
cpr3_h3_t80	one-shot	67.14 ± 24.96	33.10 ± 10.82	34.05 ± 22.73
cpr3_h3_t80	HL2	68.11 ± 25.95	45.86 ± 15.67	40.02 ± 11.98
cpr3_h3_t80	HL40	65.01 ± 32.83	29.30 ± 14.69	21.17 ± 10.88
pd3_c02_r11_neutral	one-shot	59.29 ± 32.06	18.57 ± 19.21	17.38 ± 18.25
pd3_c02_r11_neutral	HL2	54.61 ± 41.20	24.23 ± 15.56	24.11 ± 17.72
pd3_c02_r11_neutral	HL40	54.54 ± 43.41	14.11 ± 9.61	8.59 ± 9.95
pd3_c02_r14_neutral	one-shot	54.76 ± 32.05	21.19 ± 19.55	18.33 ± 19.51
pd3_c02_r14_neutral	HL2	54.87 ± 40.93	23.75 ± 15.32	23.30 ± 16.64
pd3_c02_r14_neutral	HL40	53.45 ± 42.31	14.46 ± 10.83	9.09 ± 10.48
pd3_c02_r20_neutral	one-shot	58.33 ± 33.71	16.19 ± 20.31	17.86 ± 22.72
pd3_c02_r20_neutral	HL2	53.06 ± 41.18	20.58 ± 17.98	19.58 ± 17.86
pd3_c02_r20_neutral	HL40	52.78 ± 41.35	12.29 ± 10.65	9.32 ± 10.42
pd3_c02_r24_neutral	one-shot	57.62 ± 30.95	22.38 ± 23.07	22.62 ± 23.82
pd3_c02_r24_neutral	HL2	54.38 ± 40.54	24.43 ± 14.96	23.51 ± 17.12
pd3_c02_r24_neutral	HL40	53.71 ± 42.25	15.28 ± 14.16	11.39 ± 13.30
pd3_c06_r11_neutral	one-shot	57.62 ± 30.20	20.71 ± 20.52	16.90 ± 18.34
pd3_c06_r11_neutral	HL2	54.92 ± 40.80	24.61 ± 17.10	22.55 ± 17.37
pd3_c06_r11_neutral	HL40	53.60 ± 42.64	14.60 ± 9.93	9.30 ± 12.01
pd3_c06_r14_neutral	one-shot	58.10 ± 33.66	19.29 ± 21.73	16.90 ± 21.68
pd3_c06_r14_neutral	HL2	55.06 ± 41.14	23.76 ± 15.40	22.85 ± 16.45
pd3_c06_r14_neutral	HL40	53.81 ± 43.03	14.78 ± 11.61	10.33 ± 11.39
pd3_c06_r20_neutral	one-shot	57.62 ± 32.40	16.19 ± 20.99	16.67 ± 21.17
pd3_c06_r20_neutral	HL2	53.40 ± 41.50	20.49 ± 17.14	19.15 ± 16.45
pd3_c06_r20_neutral	HL40	53.41 ± 42.24	13.50 ± 10.60	9.13 ± 10.49
pd3_c06_r24_neutral	one-shot	59.52 ± 32.54	25.00 ± 26.28	21.90 ± 23.89
pd3_c06_r24_neutral	HL2	55.70 ± 40.61	25.03 ± 16.92	24.56 ± 19.35
pd3_c06_r24_neutral	HL40	54.95 ± 43.41	17.66 ± 12.62	12.85 ± 15.17
pd3_c12_r11_neutral	one-shot	56.43 ± 33.44	18.81 ± 21.89	14.76 ± 21.27
pd3_c12_r11_neutral	HL2	54.60 ± 40.84	15.95 ± 15.54	13.99 ± 15.45
pd3_c12_r11_neutral	HL40	54.88 ± 43.13	11.39 ± 11.42	8.25 ± 11.97

Configuration	Exposure	Base	SFT	A
pd3_c12_r14_neutral	one-shot	55.95 ± 33.73	20.71 ± 25.18	16.43 ± 21.85
pd3_c12_r14_neutral	HL2	53.70 ± 40.91	15.27 ± 15.61	14.08 ± 15.37
pd3_c12_r14_neutral	HL40	53.90 ± 43.20	11.48 ± 11.81	8.63 ± 12.10
pd3_c12_r20_neutral	one-shot	58.57 ± 33.24	17.38 ± 23.47	14.29 ± 22.27
pd3_c12_r20_neutral	HL2	54.03 ± 40.82	15.51 ± 15.95	12.61 ± 15.81
pd3_c12_r20_neutral	HL40	52.65 ± 41.14	12.00 ± 12.58	8.21 ± 12.36
pd3_c12_r24_neutral	one-shot	59.76 ± 33.52	20.24 ± 25.74	22.14 ± 26.61
pd3_c12_r24_neutral	HL2	54.69 ± 40.56	17.52 ± 18.95	16.71 ± 19.36
pd3_c12_r24_neutral	HL40	54.38 ± 42.27	14.81 ± 14.79	10.73 ± 14.31
vd3_b2_c025	one-shot	78.81 ± 19.19	55.95 ± 21.41	61.19 ± 27.43
vd3_b2_c025	HL2	79.46 ± 23.40	68.53 ± 15.70	62.10 ± 23.39
vd3_b2_c025	HL40	81.15 ± 27.38	51.57 ± 16.31	43.25 ± 18.93
vd3_b2_c075	one-shot	76.19 ± 19.43	58.57 ± 20.78	61.43 ± 26.83
vd3_b2_c075	HL2	78.23 ± 24.10	67.39 ± 15.80	59.63 ± 22.13
vd3_b2_c075	HL40	76.67 ± 27.55	50.77 ± 15.59	42.49 ± 19.01
vd3_b2_c1p25	one-shot	75.71 ± 18.33	53.33 ± 22.99	55.95 ± 28.17
vd3_b2_c1p25	HL2	75.68 ± 25.27	58.19 ± 13.11	51.63 ± 20.55
vd3_b2_c1p25	HL40	72.01 ± 28.56	40.17 ± 12.40	34.41 ± 12.79
vd3_b2_c1p75	one-shot	74.29 ± 20.36	55.48 ± 22.52	53.10 ± 26.60
vd3_b2_c1p75	HL2	74.62 ± 25.56	55.45 ± 13.42	49.28 ± 19.24
vd3_b2_c1p75	HL40	70.19 ± 33.44	38.96 ± 13.68	30.60 ± 11.50
vd3_b4_c05	one-shot	78.10 ± 16.48	56.90 ± 21.37	60.95 ± 27.94
vd3_b4_c05	HL2	79.74 ± 23.51	69.87 ± 16.47	62.39 ± 21.61
vd3_b4_c05	HL40	80.39 ± 26.32	52.07 ± 17.57	47.48 ± 19.69
vd3_b4_c15	one-shot	75.00 ± 16.36	58.57 ± 19.26	58.81 ± 30.38
vd3_b4_c15	HL2	78.37 ± 24.34	62.06 ± 11.81	55.75 ± 20.77
vd3_b4_c15	HL40	76.87 ± 26.96	47.68 ± 15.94	41.99 ± 18.89
vd3_b4_c25	one-shot	74.05 ± 15.66	55.24 ± 21.29	60.71 ± 28.53
vd3_b4_c25	HL2	76.97 ± 24.95	62.89 ± 14.89	54.28 ± 21.09
vd3_b4_c25	HL40	77.54 ± 25.85	46.00 ± 14.30	35.98 ± 13.60
vd3_b4_c35	one-shot	76.19 ± 15.92	57.86 ± 21.98	56.67 ± 28.41
vd3_b4_c35	HL2	73.89 ± 26.72	57.64 ± 12.15	48.99 ± 18.45
vd3_b4_c35	HL40	67.16 ± 31.25	40.48 ± 11.52	32.19 ± 12.55
vd3_b8_c1	one-shot	73.10 ± 21.99	61.90 ± 19.57	69.52 ± 25.94
vd3_b8_c1	HL2	80.88 ± 22.27	66.30 ± 14.20	61.00 ± 22.77
vd3_b8_c1	HL40	80.12 ± 25.55	52.21 ± 18.12	47.36 ± 19.73
vd3_b8_c3	one-shot	75.24 ± 19.42	63.81 ± 19.48	68.33 ± 22.93
vd3_b8_c3	HL2	79.50 ± 23.12	64.11 ± 14.14	58.15 ± 22.28
vd3_b8_c3	HL40	76.19 ± 26.28	49.53 ± 17.42	43.18 ± 19.11
vd3_b8_c5	one-shot	75.48 ± 19.31	59.76 ± 22.04	61.90 ± 29.11
vd3_b8_c5	HL2	78.05 ± 23.74	61.38 ± 15.37	55.40 ± 22.05
vd3_b8_c5	HL40	75.77 ± 26.36	46.50 ± 13.87	39.01 ± 13.37
vd3_b8_c7	one-shot	71.67 ± 20.23	57.38 ± 23.23	58.10 ± 27.05
vd3_b8_c7	HL2	76.07 ± 25.35	60.01 ± 14.53	53.44 ± 21.73
vd3_b8_c7	HL40	73.22 ± 27.88	47.60 ± 15.13	38.63 ± 13.29
sh3_k2_h3p15	one-shot	86.43 ± 18.34	72.86 ± 23.76	72.86 ± 27.53
sh3_k2_h3p15	HL2	87.41 ± 19.95	81.91 ± 12.22	87.90 ± 9.17
sh3_k2_h3p15	HL40	87.24 ± 22.20	84.19 ± 9.70	92.08 ± 6.37
sh3_k2_h3p3	one-shot	88.33 ± 14.91	68.57 ± 25.43	74.52 ± 20.04
sh3_k2_h3p3	HL2	87.63 ± 19.54	81.73 ± 12.77	87.38 ± 10.17
sh3_k2_h3p3	HL40	87.20 ± 22.55	84.36 ± 9.30	90.01 ± 8.94
sh3_k2_h3p6	one-shot	85.95 ± 19.15	70.95 ± 26.00	75.71 ± 20.41
sh3_k2_h3p6	HL2	87.46 ± 20.26	83.95 ± 13.08	89.88 ± 7.53
sh3_k2_h3p6	HL40	86.65 ± 23.31	88.10 ± 8.36	92.55 ± 5.32
sh3_k2_h4p2	one-shot	87.14 ± 16.63	76.90 ± 21.72	79.52 ± 19.26
sh3_k2_h4p2	HL2	87.77 ± 19.75	84.33 ± 10.97	90.39 ± 7.48
sh3_k2_h4p2	HL40	87.20 ± 22.73	90.89 ± 7.22	94.12 ± 5.28
sh3_k3_h3p15	one-shot	84.76 ± 17.65	71.67 ± 26.75	68.10 ± 26.92
sh3_k3_h3p15	HL2	87.00 ± 20.25	69.23 ± 21.12	76.57 ± 14.89
sh3_k3_h3p15	HL40	87.39 ± 22.05	65.87 ± 19.57	60.66 ± 24.91
sh3_k3_h3p3	one-shot	87.62 ± 16.27	65.48 ± 25.80	68.57 ± 24.22
sh3_k3_h3p3	HL2	87.04 ± 19.87	64.74 ± 19.97	74.67 ± 20.22
sh3_k3_h3p3	HL40	86.48 ± 23.14	58.56 ± 27.17	50.74 ± 31.73

Configuration	Exposure	Base	SFT	A
sh3_k3_h3p6	one-shot	87.86 ± 18.82	70.71 ± 26.54	71.67 ± 23.92
sh3_k3_h3p6	HL2	87.02 ± 20.21	69.60 ± 18.66	78.73 ± 16.48
sh3_k3_h3p6	HL40	84.69 ± 21.19	68.59 ± 28.46	67.47 ± 17.95
sh3_k3_h4p2	one-shot	85.48 ± 17.15	70.24 ± 28.91	71.90 ± 24.62
sh3_k3_h4p2	HL2	87.18 ± 20.54	70.77 ± 21.88	82.06 ± 10.72
sh3_k3_h4p2	HL40	87.27 ± 22.30	70.72 ± 24.55	76.37 ± 17.36
td16_tau025_neutral	one-shot	31.24 ± 20.79	55.83 ± 17.55	55.29 ± 14.72
td16_tau025_neutral	HL2	46.77 ± 30.49	71.53 ± 25.09	67.66 ± 21.25
td16_tau025_neutral	HL40	47.21 ± 29.90	71.54 ± 23.35	68.46 ± 22.69
td16_tau05_neutral	one-shot	29.57 ± 20.14	57.76 ± 14.78	55.64 ± 14.13
td16_tau05_neutral	HL2	46.12 ± 30.09	71.92 ± 21.88	66.24 ± 22.88
td16_tau05_neutral	HL40	47.17 ± 27.83	72.25 ± 23.58	64.41 ± 20.02
td16_tau10_neutral	one-shot	32.62 ± 21.75	57.21 ± 18.20	51.71 ± 14.41
td16_tau10_neutral	HL2	46.75 ± 31.15	69.16 ± 28.45	65.01 ± 24.00
td16_tau10_neutral	HL40	47.91 ± 30.23	71.11 ± 23.72	64.84 ± 21.42
td16_tau20_neutral	one-shot	32.45 ± 20.45	54.31 ± 12.68	55.29 ± 13.95
td16_tau20_neutral	HL2	45.80 ± 28.70	70.18 ± 22.84	66.69 ± 20.06
td16_tau20_neutral	HL40	40.31 ± 22.63	71.27 ± 24.82	64.86 ± 21.84
td32_tau025_neutral	one-shot	24.65 ± 22.52	50.56 ± 18.28	49.72 ± 15.23
td32_tau025_neutral	HL2	35.64 ± 30.26	67.16 ± 25.26	65.08 ± 22.15
td32_tau025_neutral	HL40	36.33 ± 29.33	68.59 ± 25.68	63.13 ± 24.58
td32_tau05_neutral	one-shot	26.00 ± 18.91	49.71 ± 17.41	50.59 ± 16.07
td32_tau05_neutral	HL2	35.67 ± 31.77	66.87 ± 24.60	62.87 ± 23.75
td32_tau05_neutral	HL40	36.27 ± 31.49	67.78 ± 28.55	65.63 ± 22.43
td32_tau10_neutral	one-shot	23.58 ± 18.76	49.74 ± 17.26	50.84 ± 18.04
td32_tau10_neutral	HL2	36.89 ± 31.62	67.19 ± 26.16	64.29 ± 23.69
td32_tau10_neutral	HL40	38.02 ± 28.05	66.18 ± 26.76	63.79 ± 21.41
td32_tau20_neutral	one-shot	25.79 ± 19.12	48.48 ± 17.07	52.88 ± 18.07
td32_tau20_neutral	HL2	35.94 ± 28.69	66.83 ± 26.29	63.25 ± 24.57
td32_tau20_neutral	HL40	36.31 ± 24.52	68.88 ± 25.43	62.93 ± 21.92
td64_tau025_neutral	one-shot	16.47 ± 15.93	32.40 ± 7.60	41.34 ± 14.97
td64_tau025_neutral	HL2	31.24 ± 23.21	57.35 ± 28.32	51.82 ± 22.42
td64_tau025_neutral	HL40	39.46 ± 28.34	57.00 ± 23.95	51.88 ± 17.60
td64_tau05_neutral	one-shot	19.96 ± 18.05	34.32 ± 7.62	41.05 ± 17.41
td64_tau05_neutral	HL2	33.97 ± 22.72	57.05 ± 29.14	52.87 ± 25.26
td64_tau05_neutral	HL40	41.88 ± 30.36	53.72 ± 22.41	51.15 ± 19.30
td64_tau10_neutral	one-shot	17.62 ± 17.97	32.82 ± 6.83	41.81 ± 16.31
td64_tau10_neutral	HL2	29.33 ± 19.55	56.45 ± 29.57	51.85 ± 24.72
td64_tau10_neutral	HL40	40.58 ± 30.08	56.15 ± 26.03	48.25 ± 15.59
td64_tau20_neutral	one-shot	20.48 ± 19.03	35.40 ± 8.69	44.90 ± 19.36
td64_tau20_neutral	HL2	30.36 ± 22.68	58.21 ± 28.39	52.04 ± 24.37
td64_tau20_neutral	HL40	39.35 ± 29.33	56.10 ± 26.86	49.96 ± 18.89
td8_tau025_neutral	one-shot	25.31 ± 19.77	47.96 ± 10.57	49.59 ± 12.32
td8_tau025_neutral	HL2	36.97 ± 34.13	69.14 ± 16.81	64.96 ± 16.77
td8_tau025_neutral	HL40	35.33 ± 33.73	71.71 ± 23.03	66.89 ± 13.43
td8_tau05_neutral	one-shot	26.48 ± 23.07	45.97 ± 7.59	51.43 ± 7.26
td8_tau05_neutral	HL2	35.47 ± 32.92	67.16 ± 18.14	65.90 ± 16.19
td8_tau05_neutral	HL40	33.79 ± 32.19	73.87 ± 16.71	67.36 ± 14.00
td8_tau10_neutral	one-shot	24.85 ± 18.85	46.22 ± 9.21	46.28 ± 8.73
td8_tau10_neutral	HL2	36.35 ± 32.90	68.27 ± 17.85	64.24 ± 15.47
td8_tau10_neutral	HL40	36.09 ± 29.95	74.57 ± 17.38	68.84 ± 17.23
td8_tau20_neutral	one-shot	27.19 ± 22.27	48.67 ± 10.60	48.62 ± 13.83
td8_tau20_neutral	HL2	35.97 ± 32.27	67.85 ± 17.87	64.39 ± 16.47
td8_tau20_neutral	HL40	35.79 ± 33.20	73.39 ± 17.55	63.76 ± 14.26
tg3_m2_s02_neutral	one-shot	80.71 ± 18.00	69.05 ± 18.98	64.29 ± 23.11
tg3_m2_s02_neutral	HL2	86.84 ± 19.52	69.53 ± 19.14	68.25 ± 18.87
tg3_m2_s02_neutral	HL40	86.48 ± 21.09	62.12 ± 21.79	62.62 ± 19.68
tg3_m2_s04_neutral	one-shot	83.33 ± 15.28	68.33 ± 19.81	63.57 ± 24.05
tg3_m2_s04_neutral	HL2	86.41 ± 19.97	69.14 ± 19.01	67.87 ± 18.57
tg3_m2_s04_neutral	HL40	87.10 ± 20.81	60.91 ± 21.25	61.82 ± 21.45
tg3_m2_s06_neutral	one-shot	81.43 ± 16.95	68.10 ± 17.91	59.52 ± 22.40
tg3_m2_s06_neutral	HL2	86.78 ± 19.23	71.00 ± 19.14	68.86 ± 19.36
tg3_m2_s06_neutral	HL40	85.79 ± 21.53	63.01 ± 20.15	60.23 ± 21.25

Configuration	Exposure	Base	SFT	A
tg3_m2_s08_neutral	one-shot	78.57 ± 20.87	66.19 ± 20.36	59.29 ± 20.86
tg3_m2_s08_neutral	HL2	85.59 ± 19.25	68.55 ± 20.73	67.98 ± 20.28
tg3_m2_s08_neutral	HL40	84.29 ± 21.80	62.85 ± 22.13	62.39 ± 20.41
tg3_m3_s02_neutral	one-shot	80.95 ± 17.21	62.38 ± 24.94	65.71 ± 23.88
tg3_m3_s02_neutral	HL2	86.54 ± 19.67	65.86 ± 18.30	64.18 ± 19.13
tg3_m3_s02_neutral	HL40	86.80 ± 20.78	58.60 ± 22.81	58.00 ± 22.54
tg3_m3_s04_neutral	one-shot	81.67 ± 16.19	63.33 ± 26.58	61.43 ± 26.47
tg3_m3_s04_neutral	HL2	86.82 ± 19.82	66.90 ± 19.17	65.08 ± 19.39
tg3_m3_s04_neutral	HL40	87.16 ± 20.96	60.01 ± 20.49	60.10 ± 21.40
tg3_m3_s06_neutral	one-shot	80.95 ± 17.21	67.38 ± 20.04	62.62 ± 23.55
tg3_m3_s06_neutral	HL2	86.55 ± 19.64	67.98 ± 19.12	65.10 ± 18.68
tg3_m3_s06_neutral	HL40	85.68 ± 20.86	59.03 ± 23.22	60.11 ± 20.90
tg3_m3_s08_neutral	one-shot	81.90 ± 15.97	61.19 ± 22.70	60.24 ± 25.27
tg3_m3_s08_neutral	HL2	85.39 ± 20.80	64.84 ± 18.59	62.95 ± 19.87
tg3_m3_s08_neutral	HL40	85.06 ± 21.12	57.12 ± 21.69	58.18 ± 21.23
tg3_m4_s02_neutral	one-shot	81.19 ± 18.35	63.10 ± 17.12	62.14 ± 24.24
tg3_m4_s02_neutral	HL2	87.15 ± 19.16	67.10 ± 18.07	64.21 ± 19.14
tg3_m4_s02_neutral	HL40	87.79 ± 20.37	60.97 ± 22.18	59.14 ± 23.18
tg3_m4_s04_neutral	one-shot	80.48 ± 16.77	65.00 ± 17.95	58.81 ± 22.64
tg3_m4_s04_neutral	HL2	86.95 ± 19.05	65.44 ± 19.45	62.95 ± 19.96
tg3_m4_s04_neutral	HL40	87.17 ± 20.94	58.45 ± 21.72	60.53 ± 23.96
tg3_m4_s06_neutral	one-shot	81.67 ± 17.90	64.76 ± 20.19	57.86 ± 24.75
tg3_m4_s06_neutral	HL2	87.05 ± 19.47	67.26 ± 18.74	65.76 ± 18.87
tg3_m4_s06_neutral	HL40	87.47 ± 20.79	61.09 ± 21.35	61.34 ± 21.90
tg3_m4_s08_neutral	one-shot	82.62 ± 15.27	63.33 ± 22.32	59.76 ± 21.93
tg3_m4_s08_neutral	HL2	87.58 ± 18.83	67.02 ± 19.94	65.64 ± 20.73
tg3_m4_s08_neutral	HL40	87.87 ± 20.09	57.93 ± 22.63	60.14 ± 22.76

Table 35 keeps every complete payoff combination separate while averaging lineages. In contrast, the four-point tables in Appendix H average secondary parameters to expose the main payoff axis, with Tables 17–22 retaining each lineage separately. These are complementary views of the same canonical observations.

R Payoff Interpretation and Behavioral Comparisons

This appendix connects the observed cooperation rates to earned payoffs and the value of future reciprocity, using the game functions in Appendix C.

Why study cooperation? Across the tested configurations, the fully cooperative profile gives every player more than the fully noncooperative profile: maximum versus minimum claims in TD, investment with both returns versus no investment in Trust, and all cooperative versus all noncooperative actions elsewhere. Repeating the cooperative profile preserves that advantage. The obstacles differ: PGG, Donation-PD, CPR, TD, and Trust reward unilateral deviation; Assurance requires enough participants; Volunteer needs one provider while each player prefers another to pay. The derivations below separate these immediate incentives, collective outcomes, and future reciprocity.

R.1 What changes along each payoff axis

Let C denote the designated cooperative action, D its binary alternative, and $a_i = 1$ indicate contribution. Write u_C, u_D for the focal player's one-round payoffs from these actions. Hold others' current simultaneous actions, or their distribution where explicitly stated, fixed.

PGG and Donation-PD. For $\alpha = r/n$, PGG pays $u_i = 1 - a_i + \alpha \sum_j a_j$, hence $u_D - u_C = 1 - \alpha > 0$ with others' current actions held fixed. Define per-round group welfare as total monetary payoff, $W = \sum_i u_i = n + (r-1)\sum_i a_i$. One contribution therefore raises W by $r - 1 > 0$. Mutual contribution gives every player $r > 1$, compared with 1 under mutual retention. Across our

grid, defection remains the private stage-game optimum; Appendix R.2 separately establishes when contribution is a best response on the cooperative path to the specified history-dependent partner strategies. Donation-PD instead pays $u_i = 2 - ca_i + b\sum_{j\neq i} a_j$ with $R = 2b/c$. Its private gap is c, independent of R at fixed c, while the total-payoff increment is $c(R-1)$. The oriented plotting axes therefore summarize distinct economic margins: private contribution cost in PGG and external return in Donation-PD.

Volunteer and Assurance. For volunteering benefit B, cost $c = \rho B$, and probability π_0 that neither other player volunteers, $\mathbb{E}[u_C - u_D] = B(\pi_0 - \rho)$. This expression allows dependence between the other players' actions. In Assurance, if N counts other contributors and k includes oneself, $\mathbb{E}[u_C - u_D] = 3[(1+\lambda)\Pr(N \geq k-1) - 1]$. Contribution is weakly privately preferred under that fixed distribution when $\Pr(N \geq k-1) \geq 1/(1+\lambda)$. Changing k changes the success event. These formulas describe the monetary incentive conditional on the stated distribution of others' actions.

CPR. Actions are low extraction 1 and high extraction $h > 1$; $u_i = x_i(V - \sum_j x_j)$ with $V = 3h + 1 + 2\theta$. Writing $X = x_j + x_k \in [2, 2h]$ gives

$$u_D - u_C = (h-1)(2h + 2\theta - X) > 0 \qquad (\theta > 0). \tag{6}$$

High extraction is strictly dominant on the studied grid. Larger θ increases the private deviation advantage for a fixed opponent profile. The resource return is a stationary stage payoff, with extraction incentives reset each round.

TD. Claims are the K integers $2, \ldots, K+1$, with bonus/penalty $B_{\mathrm{TD}} = 1+\tau(K-1)$. Undercutting a common claim $z > 2$ by one yields $z - 1 + B_{\mathrm{TD}}$ instead of z, a gain $\tau(K-1)$. The minimum-claim profile remains a stage Nash equilibrium. Increasing τ raises the undercutting gain while preserving this equilibrium. The behavioral outcome is normalized claim.

From contribution rates to realized returns. Fix a lineage, stage, payoff configuration and observed trajectory b of T rounds and n players. Let $a_{ti} \in \{0, 1\}$ be contribution, u_{ti} the implemented one-round payoff, and $\hat{r}_b = (Tn)^{-1}\sum_{t,i} a_{ti}$. Define average realized player payoff per round by $\bar{u}_b = (Tn)^{-1}\sum_{t,i} u_{ti}$. Summing the PGG payoff rule gives

$$\sum_i u_{ti} = n + (r-1)\sum_i a_{ti}, \qquad \bar{u}_b = 1 + (r-1)\hat{r}_b.$$

For three-player Donation-PD, each contribution benefits two other players. With cost c, benefit b_{don} to each recipient, and return ratio $R = 2b_{\mathrm{don}}/c$, summing gives

$$\sum_i u_{ti} = 6 + (2b_{\mathrm{don}} - c)\sum_i a_{ti}, \qquad \bar{u}_b = 2 + c(R-1)\hat{r}_b.$$

The subscript on b_{don} distinguishes the donation benefit from the trajectory index. Both coefficients multiplying $\hat{r}_b$ are positive on the tested grids. Thus, at a fixed configuration, a lower contribution rate means fewer average points earned per player. This identity holds for each observed trajectory, even when players' actions depend on one another and on past rounds. At HL40, both post-trained stages have lower means than Base at all eight PGG and twelve Donation-PD configurations in Table 35. Their HL40-minus-HL2 changes are also negative in each configuration mean. Each such mean averages the seven lineages.

We first convert each run's contribution rate to points using its own payoff configuration. We then calculate stage and history differences and their means and SDs across lineages. Multiplying $\bar{u}_b$ by 200 gives average cumulative points per player over the recorded window, without discounting. This calculation describes the returns earned in the observed groups. It does not simulate what a player would earn by changing its action while holding its partners fixed.

Table 36 gives the results calculated from unrounded rates. Let $H^u_s = \bar{u}_s(40) - \bar{u}_s(2)$ be the change in mean points from HL2 to HL40, with $s \in \{B, S, A\}$. SFT earns less than Base at HL40 in 4/7 PGG lineages and 6/7 Donation-PD lineages; for Aligned, the counts are 3/7 and 6/7. Longer history lowers SFT returns in 5/7 PGG and 6/7 Donation-PD lineages, and Aligned returns in 7/7 and 5/7. These counts differ from the configuration counts because they average payoff settings within each model instead of models within each setting.

Table 36: Realized self-play payoff per player per round (points), computed from the same canonical trajectories and one-shot draws. Each entry is mean ± sample SD across seven lineages after replicate and configuration averaging. Differences and their SDs are paired within lineage. PGG uses eight configurations; Donation-PD uses twelve. These are payoff transformations of the existing cooperation observations.

Quantity	PGG	Donation-PD
Base, one-shot	2.019 ± 0.230	2.283 ± 0.159
SFT, one-shot	1.676 ± 0.123	2.096 ± 0.115
Aligned, one-shot	1.779 ± 0.128	2.091 ± 0.112
Base, HL2	1.797 ± 0.478	2.263 ± 0.197
SFT, HL2	1.594 ± 0.190	2.093 ± 0.076
Aligned, HL2	1.643 ± 0.293	2.086 ± 0.079
Base, HL40	1.763 ± 0.517	2.260 ± 0.204
SFT, HL40	1.522 ± 0.228	2.068 ± 0.061
Aligned, HL40	1.512 ± 0.373	2.049 ± 0.062
SFT−Base, HL40	-0.241 ± 0.576	-0.192 ± 0.206
Aligned−Base, HL40	-0.251 ± 0.742	-0.212 ± 0.218
H_B^u	-0.034 ± 0.057	-0.003 ± 0.010
H_S^u	-0.072 ± 0.100	-0.025 ± 0.032
H_A^u	-0.132 ± 0.102	-0.038 ± 0.042
$H_S^u - H_B^u$	-0.038 ± 0.148	-0.022 ± 0.034
$H_A^u - H_B^u$	-0.098 ± 0.150	-0.035 ± 0.046

Why the payoff mapping is game-specific. The other games connect behavior to returns through the joint action profile. In Assurance, full participation yields the highest group payoff and is a stage-game equilibrium, but an unsuccessful contributor earns zero: moving from no contributors to one lowers total payoff from 9 to 6 before the threshold is reached. In CPR, write $X = \sum_i x_i$ for total extraction; group payoff is $W = X(V - X)$. Mutual restraint benefits everyone relative to mutual high extraction, while intermediate profiles can yield still higher group totals. For example, at $h = 1.5, \theta = 0.8$, two restrained extractors give group payoff 12.6, compared with 12.3 for three. In TD, group payoff is twice the lower claim, so the average normalized claim alone does not determine it. In Trust, investment creates the group surplus; trustee returns distribute it between players, while unexecuted return intentions create no current surplus. These distinctions identify which joint outcomes must accompany the behavioral rate when evaluating group performance.

Volunteer: willingness to provide and efficient provision. If v of the three players volunteer, group payoff is $W(v) = 3B\mathbf{1}\{v \geq 1\} - vc$. Nobody volunteering gives zero to every player. One volunteer supplies the benefit to all three and earns $B - c > 0$ itself; further volunteers incur additional costs without increasing provision. With $B = 4, c = 1.5$, group payoff is 0 with no volunteer, 10.5 with one, and 7.5 with three. The relevant cooperative achievement is reliable provision with an efficient allocation of its cost. The cooperation rate measures players' willingness to supply this necessary benefit, while provision frequency measures whether someone actually does so. If $\bar{q} = \mathbb{E}[v]/3$ is the all-seat cooperation rate, then $\mathbb{E}[W] = 3B \Pr(v \geq 1) - 3c\bar{q}$, with no independence assumption. A lower rate can therefore reflect either successful cost sharing or more failures to provide; the joint outcome distinguishes them. This is why Volunteer belongs in the behavioral audit, with provision and redundancy as its natural welfare interpretation (Appendix C.4).

R.2 An elementary Nash-reversion benchmark

This benchmark states when future mutual gains outweigh an immediate deviation gain under a specified cooperative strategy. Section 5.2 separately measures responses to instructions about future interaction and compares them with human behavior.

Consider an idealized infinite repetition of a stationary finite-action game with bounded payoffs, perfect public monitoring, no implementation noise, and a common discount factor $\delta \in [0, 1)$ applied to expected monetary payoff. Suppose a symmetric cooperative profile gives each player v_C, the best

single-period deviation against it gives v_D, and an executable symmetric stage-Nash profile gives v_N, with $v_D > v_C > v_N$. Prescribe cooperation until any deviation, then the stage-Nash profile forever. Its cooperative-path incentive condition is

$$\frac{v_C}{1-\delta} \geq v_D + \frac{\delta v_N}{1-\delta} \quad \Longleftrightarrow \quad \delta \geq \delta^* := \frac{v_D - v_C}{v_D - v_N}. \tag{7}$$

Proof. The left side is the value of compliance and the right side is the best immediate deviation followed by the prescribed punishment. Rearrangement gives the threshold. During punishment, no one-period deviation improves the current payoff, since the prescribed profile is stage Nash, and punishment is absorbing. The one-deviation principle for discounted finite-action repeated games then establishes sequential incentives after every history. If the inequality fails, a cooperative-path deviation is profitable. The condition is therefore necessary and sufficient for this *specified strategy profile*. Playing stage Nash forever is also an equilibrium at every δ.

For the implemented payoff rules, the following expressions follow by substitution. Each table entry is a deterministic payoff identity.

Game	v_C	v_D	v_N	δ^*
PGG	$n\alpha$	$1 + (n-1)\alpha$	1	$\frac{1-\alpha}{(n-1)\alpha}$
Donation-PD	$2 + c(R-1)$	$2 + cR$	2	$\frac{1}{R}$
CPR	$3h - 2 + 2\theta$	$h(2h - 1 + 2\theta)$	$h(1 + 2\theta)$	$\frac{h-1+\theta}{h}$
TD	$K + 1$	$K + B_{\mathrm{TD}}$	2	$\frac{\tau}{1+\tau}$

For PGG the punishment is all-retain, for Donation-PD all-no-donation, for CPR all-high, and for TD all-minimum. Here $n\alpha > 1$, $c > 0$, $R > 1$, and $0 < \theta < 1$ ensure the required payoff orderings; in CPR, $v_C - v_N = 2(h-1)(1-\theta) > 0$. In TD, undercutting the maximum by one is the best deviation, while a higher claim against the minimum receives $2 - B_{\mathrm{TD}} < 2$. Fudenberg & Maskin (1986) provide the broader repeated-game equilibrium background. The benchmark assumes perfect monitoring, credible punishment, monetary preferences, infinite repetition, and the specified strategy. Cooperative and stage-Nash behavior can both be supported. The CONTINUE prompt instead specifies a stated 99% continuation chance, which is distinct from an internal discount factor.

S Detailed Related Work

S.1 Post-training and social-agent evaluation

Instruction tuning and learning from human feedback improve evaluated instruction following and preferences (Ouyang et al., 2022); DPO optimizes preferences without the same explicit reward-model-and-RL pipeline (Rafailov et al., 2023). Generalization comparisons of SFT and RL (Chu et al., 2025), and the distinction between high-probability reasoning success and solutions recoverable by sampling (Yue et al., 2025), emphasize defining the evaluation target. We compare released stages on their own generated cooperation, payoff responses, and history responses across several preference- and reward-optimization objectives.

Economic-game evaluations already demonstrate sensitivity to objectives, prompts, partners and repetition (Phelps & Russell, 2025; Fontana et al., 2025; Akata et al., 2025). SOTOPIA evaluates interactive goals and relationships (Zhou et al., 2024); GovSim adds resource sustainability over time (Piatti et al., 2024), unlike our stationary CPR payoff. Social training studies investigate behavior cloning, self-reinforcement, reward design, and incentive-supported coordination (Wang et al., 2024; Yu et al., 2025; Liang et al., 2025). CoopEval compares repetition, reputation, mediation and contracting (Tewolde et al., 2026). We complement these studies by comparing released training stages under common game rules and evaluation procedures.

S.2 Memory, conditional cooperation, and recovery

Liu et al. (2026) introduce the memory curse as lower cooperation with more visible interaction history and study historical content, reasoning, and interventions. We ask how this decline changes along a training sequence. Our comparisons $K_{S/B}$ and $K_{A/B}$ measure whether longer histories reduce cooperation more after post-training than at Base; $K_{A/S}$ compares SFT with its later Aligned checkpoint. We reuse part of the experimental framework but no data or results (Appendix M). Studies of reasoning in public goods and norm enforcement further motivate examining both the rationale and the selected action (Piedrahita et al., 2025; Li & Shirado, 2025).

Human conditional-contribution schedules (Fischbacher et al., 2001), separately elicited preferences and beliefs (Fischbacher & Gächter, 2010), and distributional/reciprocity choices (Charness & Rabin, 2002) motivate separating an opening action from its subsequent feedback. Ma et al. (2021) display zero to four encounters within anonymous eight-person groups and find an intermediate-history peak; we instead compare caps of two and forty in fixed groups. Noisy repeated-PD experiments motivate studying tolerance and recovery after temporary defection (Fudenberg et al., 2012). Our generated self-play and fixed-record probes resolve different parts of that question: the former changes all players and their resulting histories; the latter compares responses to a shared input.

S.3 Game-specific payoff responses and human comparisons

Human public-goods studies separately manipulate MPCR/group size in ten-period play (Isaac & Walker, 1988) and private/external returns in ten no-feedback one-shot decisions (Goeree et al., 2002); a genuine single-decision study reports no statistically detected MPCR response (Struwe et al., 2024). These designs motivate identifying both the economic margin and the interaction protocol. Sequential dilemmas vary a second mover's private deviation gain and harm to a cooperative first mover (Baader et al., 2026). Their positive mutual-cooperation reference supports relative-gain measures; our response statistic instead uses the tested span of each game's raw parameter.

Human TD claims fall with larger reward/penalty sizes under random rematching (Capra et al., 1999); related LLM signatures are studied by Roberts et al. (2025). Human Trust experiments elicit investment and returns in a sequential two-person design (Berg et al., 1995); ours elicits one investor's and two trustees' simultaneous contingent intentions. Behavioral-test similarity (Mei et al., 2024) and prediction of observed human choices (Shapira et al., 2026; Binz et al., 2026) are different evaluation targets; the latter also includes one-shot/first-round settings favoring aligned models. Our comparisons concern specific payoff and history responses under each study's stated protocol.

S.4 Continuation, presentation, and shared-input comparisons

Repeated-game theory distinguishes equilibrium support from equilibrium selection (Fudenberg & Maskin, 1986); Appendix R.2 gives the specified Nash-reversion benchmark. Human repeated-PD evidence studies this distinction empirically (Dal Bó & Fréchette, 2011). Public-goods studies use different finite or probabilistic stopping procedures and report different cooperation patterns (Lugovskyy et al., 2017; Anderson et al., 2024). They motivate specifying both what agents are told about future rounds and how collection actually stops. Our canonical data record 200 rounds under continuing instructions; the shared-record tests vary a current decision's instructions at fixed presented history and payoff.

Externality-framing experiments change how a decision is presented while keeping its potential payoffs fixed (Andreoni, 1995). Our PGG comparison shows or hides past records, which also changes prompt length. In the final-round condition, contributing still changes the player's current payoff by $r/n - 1$, regardless of the records shown. These tests ask how a model responds to the information it receives, alongside our measurements of cooperation during actual interaction.

T A Seven-Stage Nemotron-Cascade Case Study

T.1 Checkpoints, payoff settings, and measurements

This case study follows one released lineage through seven checkpoints: Qwen3-8B-Base, SFT, RLHF, instruction-following RL (IFRL), mathematical RL (MathRL), code RL (CodeRL), and

software-engineering RL (SWE-RL). The canonical Aligned checkpoint for this lineage is *RLHF*. SWE-RL is the endpoint of this longer case study. The four additional checkpoints extend the case without changing the primary 21-checkpoint panel.

Base uses `Qwen/Qwen3-8B-Base`. SFT through CodeRL use the correspondingly named subfolders of `nvidia/Nemotron-Cascade-8B-Intermediate-ckpts`; SWE-RL uses `nvidia/Nemotron-Cascade-8B`. All checkpoints use the same prompt template and action-label interface, bfloat16 inference, temperature 0.7, top-p 0.8, top-k 20, a 256-token rationale cap and an 8-token constrained action cap.

All seven checkpoints are evaluated on the same original 82 configurations: PGG uses five MPCR values, CPR three θ values, and the other grids are unchanged. Each checkpoint has 20 one-shot draw IDs per configuration and five trajectory seeds at each of HL2 and HL40; repeated rates average all 200 logged rounds and all seats within configuration. This common-grid case and the amended primary panel have the distinct scopes summarized in Appendix E.

We compare cooperation rates and payoff responses on the same five games: PGG, CPR, Donation-PD, Volunteer, and Assurance. A separate history table adds Trust to match the six-binary-game average in the main text. For each checkpoint and run, we first average cooperation over configurations within a game, then average games equally. For payoff responses, we fit the main parameter while holding secondary parameters fixed, then average these fits and games (Equation 2). PGG therefore fits five points here and CPR three. Assurance uses the λ span 0.35; fitting successful payoff $h = 3(1 + \lambda)$ over span 1.05 gives the identical fitted drop.

The tables report mean $\pm$ sample SD across 20 one-shot draws or five repeated-game runs *within a checkpoint*. For history and consecutive-stage comparisons, we subtract results with the same seed ID before computing the SD. This SD describes variation across runs of this one lineage. Rows follow the released training sequence.

T.2 Levels and responsiveness need not move together

Figure 5 and Tables 37–38 compare the cooperation rate and fitted payoff drop across these same five games. The RLHF-to-IFRL transition increases mean cooperation by 11.22 ± 4.63, 6.18 ± 1.44, and 14.53 ± 3.64 pp in OS, HL2 and HL40, while fitted responsiveness changes by -6.87 ± 9.39, -1.51 ± 3.36, and -11.88 ± 7.27 pp, respectively. Thus instruction-following RL raises cooperation but makes its average response to payoff changes smaller in all three conditions. The figure shows how the two measurements change over training; the tables give each checkpoint's mean, SD, and history response.

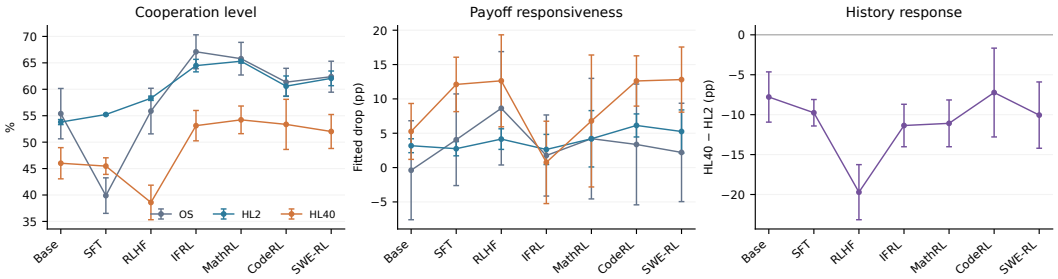

Figure 5: The same five symmetric binary games across seven checkpoints: cooperation level, all-point fitted payoff responsiveness, and HL40–HL2 history response. Points and whiskers show mean $\pm$ within-checkpoint replicate SD (20 one-shot draws or five trajectory seeds); the history contrast pairs the five seeds. Each panel has its own labeled metric and axis. RLHF is the canonical Aligned endpoint.

Table 37: Same five symmetric binary games for both level and response: PGG, CPR, Donation-PD, Volunteer, Assurance. Levels are percentages; history contrasts are pp. Mean ± within-checkpoint replicate SD, not seven-lineage SD. Repeated contrasts pair five nominal seeds.

Stage	OS	HL2	HL40	H	$K_{s/B}$
Base	55.39 ± 4.76	53.80 ± 0.48	46.01 ± 2.95	-7.79 ± 3.15	$+0.00 \pm 0.00$
SFT	39.89 ± 3.36	55.23 ± 0.16	45.47 ± 1.57	-9.76 ± 1.66	-1.97 ± 4.53
RLHF	55.88 ± 4.31	58.30 ± 0.41	38.59 ± 3.27	-19.71 ± 3.45	-11.92 ± 5.32
IFRL	67.10 ± 3.21	64.48 ± 1.18	53.12 ± 2.87	-11.36 ± 2.66	-3.57 ± 4.61
MathRL	65.79 ± 3.09	65.31 ± 0.35	54.22 ± 2.62	-11.09 ± 2.93	-3.30 ± 1.97
CodeRL	61.35 ± 2.62	60.59 ± 1.93	53.36 ± 4.74	-7.23 ± 5.56	$+0.56 \pm 6.12$
SWE-RL	62.38 ± 2.93	62.07 ± 1.39	52.01 ± 3.21	-10.05 ± 4.15	-2.27 ± 2.96

Table 38: Same five symmetric binary games for both level and response: PGG, CPR, Donation-PD, Volunteer, Assurance. All-point fitted responses in pp; mean ± replicate SD (20 draws for OS, five seeds for HL2/HL40).

Stage	OS	HL2	HL40
Base	-0.39 ± 7.21	$+3.18 \pm 1.01$	$+5.26 \pm 4.07$
SFT	$+4.05 \pm 6.67$	$+2.75 \pm 1.04$	$+12.11 \pm 3.96$
RLHF	$+8.63 \pm 8.25$	$+4.15 \pm 1.50$	$+12.63 \pm 6.70$
IFRL	$+1.76 \pm 5.90$	$+2.64 \pm 2.19$	$+0.75 \pm 6.00$
MathRL	$+4.21 \pm 8.78$	$+4.19 \pm 4.10$	$+6.79 \pm 9.61$
CodeRL	$+3.37 \pm 8.79$	$+6.13 \pm 1.68$	$+12.61 \pm 3.66$
SWE-RL	$+2.21 \pm 7.16$	$+5.25 \pm 3.14$	$+12.81 \pm 4.74$

T.3 History responses across game sets

On the six-binary-game panel, every post-training checkpoint has a more negative mean history response than Base (Table 39). The Base and SWE-RL HL40 levels are $54.46 \pm 2.46\%$ and $54.27 \pm 2.36\%$, while their history changes are -6.43 ± 2.60 and -10.43 ± 3.04 pp. Similar endpoint levels thus coexist with different responses to history. Intermediate HL40 levels range from RLHF's 38.73% to MathRL's 57.50%.

The five-game panel supplies the like-for-like level/response comparison above. In that panel, five of six post-training history contrasts are more negative than Base; CodeRL has $K_{s/B} = +0.56 \pm 6.12$ pp. The six-binary-game panel, which includes Trust, has negative Base-relative history contrasts at all six post-training checkpoints. Both summaries specify their game sets.

Table 39: Six-binary-game cooperation and history panel, including Trust and excluding TD. Levels are percentages; history contrasts are pp. Mean ± within-checkpoint replicate SD, not seven-lineage SD. Repeated contrasts pair five nominal seeds.

Stage	OS	HL2	HL40	H	$K_{s/B}$
Base	57.82 ± 4.91	60.89 ± 0.45	54.46 ± 2.46	-6.43 ± 2.60	$+0.00 \pm 0.00$
SFT	38.96 ± 3.60	56.44 ± 0.28	45.84 ± 1.51	-10.60 ± 1.56	-4.17 ± 3.84
RLHF	50.59 ± 3.83	58.58 ± 0.50	38.73 ± 2.81	-19.85 ± 3.02	-13.42 ± 4.61
IFRL	66.08 ± 3.42	67.63 ± 0.72	55.92 ± 2.06	-11.70 ± 1.90	-5.27 ± 3.50
MathRL	63.39 ± 3.04	69.10 ± 0.42	57.50 ± 2.19	-11.60 ± 2.26	-5.17 ± 2.05
CodeRL	59.92 ± 2.79	63.06 ± 1.05	55.53 ± 4.01	-7.53 ± 4.58	-1.09 ± 5.12
SWE-RL	60.36 ± 2.88	64.70 ± 1.02	54.27 ± 2.36	-10.43 ± 3.04	-4.00 ± 1.95

U A Controlled Case Study of Individual Action Value

This experiment asks how many expected points a model gives up through its current choice when it knows how its partners behave. It covers PGG, Donation-PD, and Assurance, all seven lineages, and

all three stages at both history lengths. The partners follow disclosed rules rather than language-model policies. We compare the model's chosen action with the best current action, assuming optimal choices afterward.

Neutral-prompt design. The 21 checkpoints face 16 payoff configurations: four PGG settings ($n = 3$, MPCR $0.4, 0.5, 0.6, 0.8$); four Donation-PD settings ($n = 3$, $c = 0.6$, $R = 1.1, 1.4, 2.0, 2.4$); and eight Assurance settings ($k = 2, 3$ crossed with $\lambda = 0.05, 0.10, 0.20, 0.40$). At each setting, partners either respond reciprocally (R), cooperate with a constant high probability (H), or cooperate with a constant low probability (L); their exact rules are given below. We test both previous-action states, one or forty remaining decisions ($m = 1, 40$), and two or forty visible past rounds ($L = 2, 40$).

Two history blocks lead to the same current state and end with the same two rounds. Earlier in the record, the focal player's cooperation frequency is either 0.8 or 0.2. Each distinct input receives eight rationale/action draws. We average configurations within game, the three rules, the two states, and the two history blocks equally, then weight games and lineages equally. If two conditions generate exactly the same input, they reuse the same query rather than count it as an independent observation.

The neutral footer N restates player count, seat, round, and output format without a state cue or an optimal-action value. The full payoff rules, partner rules, state and remaining horizon are disclosed. Serving uses temperature 0.7, top-p 0.8, top-k 20, rationale budget 256, action budget eight, and at most three attempts with no fallback action. The configurations are listed above; the partner rules, payoffs, and complete scoring calculation follow below.

Table 40: Controlled case study: current-action loss under the specified partner rules, using neutral prompts and one current choice with 39 optimally chosen future actions, discount 0.99. Lower is better. L is the visible-history cap. Entries are mean ± seven-lineage SD (points), rounded to two decimals; all allowed completions of five invalid outputs yield these same summaries. Table 41 retains the finer-precision bounds.

Game	L	Base	SFT	Aligned
PGG	2	0.26 ± 0.02	0.23 ± 0.03	0.23 ± 0.05
	40	0.25 ± 0.02	0.22 ± 0.03	0.21 ± 0.05
Donation-PD	2	0.29 ± 0.10	0.19 ± 0.09	0.17 ± 0.08
	40	0.28 ± 0.09	0.18 ± 0.09	0.16 ± 0.08
Assurance	2	0.92 ± 0.16	0.76 ± 0.21	0.72 ± 0.24
	40	0.92 ± 0.17	0.71 ± 0.23	0.67 ± 0.22

U.1 Disclosed partner rules and action values

The focal language-model player faces two programmed partners. Its action is $a \in \{0, 1\}$: cooperation is 1 and the alternative is 0. Let ξ identify the game and payoff configuration, $\kappa \in \{\mathsf{R}, \mathsf{H}, \mathsf{L}\}$ the partner rule, and $z \in \{0, 1\}$ the focal player's previous action. The two time variables refer to different things: $L = 40$ means forty past rounds are visible; $m = 40$ means the current decision and 39 future decisions are scored. The 200-round self-play experiment is separate.

Under the reciprocal rule R, each partner cooperates with probability 0.9 if the focal player cooperated last round and 0.1 otherwise: $p_{\mathsf{R}}(z) = 0.1 + 0.8z$. Under H and L, the probability stays at 0.9 or 0.1. Conditional on z, the partners choose independently, giving joint probability $P_\kappa(y \mid z) = \prod_{j=1}^{2} p_\kappa(z)^{y_j}[1 - p_\kappa(z)]^{1-y_j}$ for $y = (y_1, y_2) \in \{0, 1\}^2$. Current actions are simultaneous; the focal action a becomes the next state $z' = a$. The prompt tells the model the payoffs, partner rules, state, remaining decisions, and discount factor $\gamma = 0.99$. The objective is to maximize its own expected discounted points.

The three rules distinguish responsiveness from baseline cooperation. Rule R provides an action-dependent opportunity for future reciprocity; H and L hold cooperation probabilities high or low without responding to the focal action. The case study therefore tests choices across specified reciprocal and nonresponsive environments. These probabilities are design parameters, not fitted checkpoint policies.

Per-round payoffs. Let $c > 0$ be the contribution cost in Donation-PD and let R be its return parameter. The focal payoffs used for the reference calculation are

$$u_{\mathrm{PGG}}(a, y) = 1 - a + \alpha(a + y_1 + y_2), \tag{8}$$

$$u_{\mathrm{Donation}}(a, y) = 2 - ca + \frac{Rc}{2}(y_1 + y_2), \tag{9}$$

$$u_{\mathrm{Assurance}}(a, y) = 3(1 - a) + 3(1 + \lambda)a\,\mathbf{1}\{a + y_1 + y_2 \geq k\}. \tag{10}$$

Here α is MPCR, k includes the focal player in the success threshold, and λ is the success premium. Configurations and numerical representations are taken from the executed game factory. The reference uses the same payoffs and joint partner sampler as the model evaluation.

Backward recursion. For fixed (ξ, κ), let $V_m(z)$ be the maximum expected discounted individual return obtainable from state z with m decisions remaining, and let $Q_m(z, a)$ be that return with the current action fixed to a and all subsequent actions chosen optimally. Define

$$V_0(z) = 0, \tag{11}$$

$$Q_m(z, a) = \sum_{y \in \{0,1\}^2} P_\kappa(y \mid z)\left[u_\xi(a, y) + \gamma V_{m-1}(a)\right], \tag{12}$$

$$V_m(z) = \max_{a \in \{0,1\}} Q_m(z, a). \tag{13}$$

Each recursion step sums all four partner-action outcomes for two public states and two focal actions. Starting from $V_0 = 0$, it computes exact-model values through horizon 40; sampled language-model decisions are scored against these values. Here z is the previous focal action, distinct from the normalized payoff coordinate in Appendix A.1.

Why the recorded state is sufficient for this reference. Histories with the same (ξ, κ, z, m) have identical optimal action values. At $m = 0$ the values are zero. Inductively, equal states give the same partner distribution and payoffs, and each action gives the same next state $z' = a$ and continuation value. Equations 12–13 establish the claim. Model choices may still depend on additional historical text.

Current-action loss and oracle labels. For the action emitted by the model, define

$$\ell_m(z, a) = V_m(z) - Q_m(z, a), \qquad d_m(z) = Q_m(z, 1) - Q_m(z, 0). \tag{14}$$

The loss measures the opportunity cost of the current action with optimal continuation from the next decision onward. Oracle-C denotes $d_m(z) > 0$, oracle-D denotes $d_m(z) < 0$, and an oracle tie denotes exact equality in the rational-arithmetic reference. No tolerance or small-gap filter is applied. These are input-state categories, not extra models or experimental arms. The two binary-action losses are

$$\ell_m(z, 1) = \max\{0, -d_m(z)\}, \qquad \ell_m(z, 0) = \max\{0, d_m(z)\}. \tag{15}$$

For model cooperation probability $p(1 \mid u)$ at input u, expected loss is $\max\{0, d_m(z)\} - d_m(z)p(1 \mid u)$.

When does cooperation improve long-term individual return? The decisive question is whether today's contribution changes tomorrow's partners. Under the nonresponsive rules H and L, contributing today does not increase their future cooperation: their probabilities stay at 0.9 or 0.1 regardless of the focal action. Consequently, $V_{m-1}(1) = V_{m-1}(0)$ and the continuation terms cancel. For any remaining horizon, $Q_m(z, 1) - Q_m(z, 0) = -(1 - \alpha)$ in PGG and $-c$ in Donation-PD. Retaining the contribution therefore yields higher individual value even with 39 future decisions remaining. Repetition creates an opportunity for reciprocity, while the partner rule determines whether a current contribution actually earns that future benefit.

Under the responsive rule R, current cooperation increases each partner's next-round cooperation probability from 0.1 to 0.9. For example, with PGG $\alpha = 0.6$ and two decisions remaining, the current contribution costs 0.4 points, but its future benefit is $0.99 \times 0.6 \times 2(0.9 - 0.1)$. The terminal best action is retention, and the total current-action value difference is $-0.4 + 0.9504 = 0.5504$: cooperation is now the higher-value individual choice. This deterministic illustration shows how the same game can reward cooperation or retention depending on the opportunity for future reciprocity. The recursion accounts for that opportunity at every remaining horizon.

Normalization and missing outcomes. To average losses across games, we divide by the range of one-round payoffs at each configuration: $W_\xi = \max_{a,y} u_\xi(a, y) - \min_{a,y} u_\xi(a, y)$, using all eight possible joint actions. The normalized loss is therefore measured relative to that payoff range, not in percentage points. It can exceed one because the loss includes future decisions while the denominator covers one round. We do not clip the ratio. For an invalid decision, we keep the possible losses from both legal actions.

Stage levels. Both post-trained stages have lower mean loss than Base in all three games at both history lengths for $m = 40$ (Table 40). Table 41 also retains the terminal $m = 1$ conditions.

History changes in action loss. Write $\mathcal{L}_s^{m,L,N}$ for mean current-action loss at stage s, horizon m, history cap L, and neutral footer N. Define $H_s^{\text{loss}} = \mathcal{L}_s^{40,40,N} - \mathcal{L}_s^{40,2,N}$ and $K_{s/B}^{\text{loss}} = H_s^{\text{loss}} - H_B^{\text{loss}}$. These contrasts quantify how showing more history changes action loss at each stage and relative to Base. Table 42 reports the Base-relative intervals, which span both negative and positive changes; Table 43 gives the stage-specific history effects and intervals.

Table 41: Expected points lost through the current action, using neutral prompts, for every game, stage, remaining horizon, and history length. Entries are mean ± SD across seven lineages. Where missing outputs prevent an exact summary, superscript b marks the range obtained by assigning either legal action to them, rounded outward. This range is a missing-output bound, not a confidence interval.

Game / m / stage	$L2, N$	$L40, N$
PGG / 1 / Base	$+0.316 \pm 0.081$	$[0.305, 0.306]^{\text{b}}$
PGG / 1 / SFT	$+0.258 \pm 0.076$	$+0.223 \pm 0.072$
PGG / 1 / Aligned	$+0.257 \pm 0.070$	$+0.233 \pm 0.076$
PGG / 40 / Base	$+0.264 \pm 0.024$	$+0.249 \pm 0.022$
PGG / 40 / SFT	$+0.228 \pm 0.028$	$+0.222 \pm 0.027$
PGG / 40 / Aligned	$+0.228 \pm 0.053$	$+0.213 \pm 0.052$
Donation-PD / 1 / Base	$+0.358 \pm 0.156$	$+0.358 \pm 0.163$
Donation-PD / 1 / SFT	$+0.201 \pm 0.158$	$+0.193 \pm 0.160$
Donation-PD / 1 / Aligned	$+0.178 \pm 0.138$	$+0.167 \pm 0.141$
Donation-PD / 40 / Base	$+0.288 \pm 0.095$	$[0.275, 0.276]^{\text{b}}$
Donation-PD / 40 / SFT	$+0.187 \pm 0.090$	$+0.179 \pm 0.086$
Donation-PD / 40 / Aligned	$+0.173 \pm 0.078$	$+0.161 \pm 0.079$
Assurance / 1 / Base	$+1.061 \pm 0.220$	$[1.046, 1.048]^{\text{b}}$
Assurance / 1 / SFT	$+0.819 \pm 0.297$	$+0.765 \pm 0.291$
Assurance / 1 / Aligned	$+0.831 \pm 0.272$	$+0.750 \pm 0.279$
Assurance / 40 / Base	$+0.915 \pm 0.157$	$[0.917, 0.918]^{\text{b}}$
Assurance / 40 / SFT	$+0.756 \pm 0.206$	$+0.713 \pm 0.229$
Assurance / 40 / Aligned	$+0.723 \pm 0.240$	$+0.668 \pm 0.221$

Table 42: Does longer history change action loss more after post-training than at Base? With forty decisions remaining, $K^{\text{loss}}_{s/B} = H^{\text{loss}}_{s} - H^{\text{loss}}_{B}$ compares the HL40-minus-HL2 loss changes. Entries give mean ± SD across seven lineages and intervals adjusted for four comparisons ($M = 4$); b marks missing-output bounds. Game rows use points; the equal-game row first divides losses by the configuration's one-round payoff range.

Game / original contrast	Estimate / dispersion	Family-adjusted interval
PGG / $K^{\text{loss}}_{S/B}$	$+0.0095 \pm 0.0419$	$[-0.0196, 0.0385]$
PGG / $K^{\text{loss}}_{A/B}$	-0.0007 ± 0.0232	$[-0.0298, 0.0284]$
Donation-PD / $K^{\text{loss}}_{S/B}$	$[0.0040, 0.0043]^{\text{b}}$	$[-0.0258, 0.0342]$
Donation-PD / $K^{\text{loss}}_{A/B}$	$[-0.0002, 0.0001]^{\text{b}}$	$[-0.0301, 0.0299]$
Assurance / $K^{\text{loss}}_{S/B}$	$[-0.0458, -0.0456]^{\text{b}}$	$[-0.1288, 0.0374]$
Assurance / $K^{\text{loss}}_{A/B}$	$[-0.0575, -0.0574]^{\text{b}}$	$[-0.1406, 0.0256]$
Equal-game / span / $K^{\text{loss}}_{S/B}$	$[-0.0016, -0.0015]^{\text{b}}$	$[-0.0133, 0.0102]$
Equal-game / span / $K^{\text{loss}}_{A/B}$	$[-0.0058, -0.0057]^{\text{b}}$	$[-0.0175, 0.0060]$

Table 43: How showing forty rather than two past rounds changes current-action loss, with forty decisions remaining. $H^{\text{loss}}_{s} > 0$ means greater loss with longer history. Entries give mean ± SD across seven lineages and intervals adjusted for the original four comparisons ($M = 4$). Superscript b marks missing-output bounds. Game rows use points; the equal-game row divides each loss by its configuration's one-round payoff range before averaging.

Game / stage	H^{loss}_{s}	Existing interval
PGG / Base	-0.0151 ± 0.0245	$[-0.0356, +0.0055]$
PGG / SFT	-0.0056 ± 0.0252	$[-0.0261, +0.0150]$
PGG / Aligned	-0.0158 ± 0.0177	$[-0.0363, +0.0048]$
Donation-PD / Base	$[-0.0123, -0.0120]^{\text{b}}$	$[-0.0334, +0.0091]$
Donation-PD / SFT	-0.0080 ± 0.0230	$[-0.0291, +0.0132]$
Donation-PD / Aligned	-0.0123 ± 0.0207	$[-0.0334, +0.0089]$
Assurance / Base	$[+0.0024, +0.0025]^{\text{b}}$	$[-0.0563, +0.0612]$
Assurance / SFT	-0.0433 ± 0.0470	$[-0.1020, +0.0155]$
Assurance / Aligned	-0.0550 ± 0.0427	$[-0.1138, +0.0037]$
Equal-game / span / Base	$[-0.0056, -0.0055]^{\text{b}}$	$[-0.0138, +0.0028]$
Equal-game / span / SFT	-0.0071 ± 0.0099	$[-0.0154, +0.0012]$
Equal-game / span / Aligned	-0.0113 ± 0.0075	$[-0.0196, -0.0030]$

Sampling, aliases, and invalid outputs. When the same query appears in several conditions, we combine its weights before calculating uncertainty. For independent distinct draws, let a_j be the combined weight of draw j and $[l_j, u_j]$ its possible loss range. The weighted Hoeffding radius is $\epsilon = \sqrt{\frac{1}{2}\log(2M/\eta)\sum_j a_j^2(u_j - l_j)^2}$, with $\eta = 0.05$. We keep the original $M = 4$ adjustment for two history and two state-cue comparisons, even though this section displays only the neutral-prompt results. For invalid outputs, we allow both legal actions and add ϵ to the resulting worst-case bounds. SDs across lineages still use seven estimates.

At $m = 40, L = 40$, Donation-PD has two invalid neutral outputs among 5,376 distinct draws and Assurance three among 10,752, all at Base. Assigning either legal action to each missing output leaves the two-decimal means and SDs in Table 40 unchanged. Table 41 gives finer-precision bounds.