SELF-PLAY EVALUATION UNDER POPULATION EXPOSURE MISMATCH: UNIDENTIFIED VERDICTS ON DEFENSIVE SKILL IN COMPETITIVE MAHJONG

Anonymous authors
Paper under double-blind review

ABSTRACT

Game-playing agents are usually selected in-house, against copies of themselves or ancestors, before they meet their deployment population. We study one way this selection can mislead: a skill may pay off only in states that opponents create, and the in-house population may create those states at a different rate. We measure this population-exposure mismatch for a Chinese Standard Mahjong agent, using deal-in incidence as a policy-dependent proxy for threat-state exposure, and ask what an in-house verdict on a defensive intervention identifies. With our simulator's scoring defect corrected, our strongest in-house population exposes the deployed policy at 0.57 times the live-field rate (95% CI 0.44–0.74) and at a competition final's rate. Under stated assumptions for transporting a verdict, any exposure gap leaves the deployed sign of a fold policy with a precise negative in-house verdict unidentified: against the live field the sign flips if an exposed game gains over 0.039 placement points; against the final the verdict carries over. Gate verdicts predate the correction. A live A/B resolves neither the deal-in effect nor placement, which would take thousands of games per arm. In that final, the top two agents' statistical tie splits into opposing score channels; finalists face threat states equally often but differ in how often a threatened discard deals in. We distil an evaluation recipe that separates exposure-validity checks from general hygiene. All evidence comes from one game and one agent lineage.

1 INTRODUCTION

Agents for competitive games are selected in-house before they are deployed. A candidate plays copies of itself, its training ancestors or a curated population, and the resulting win rate or placement decides what ships (Silver et al., 2018; Vinyals et al., 2019; Li et al., 2020). The known failure modes of this practice concern how outcomes are aggregated: ratings are non-transitive (Balduzzi et al., 2018; Omidshafiei et al., 2019), strategy spaces cycle (Czarnecki et al., 2020), and naive self-play overfits its own population, which motivates population-based and league training (Lanctot et al., 2017; Vinyals et al., 2019; Jiang et al., 2024). In cooperative settings, the gap between evaluation partners and deployment-time partners is now an explicit evaluation problem (Bard et al., 2020; Carroll et al., 2019; Wang et al., 2024; Liang et al., 2024).

This paper studies a related gap in a competitive game, at the level of *states* rather than partners or ratings. Some skills pay off only in states that an opponent must create. In Mahjong the costliest single event is *dealing in*, discarding the tile an opponent claims to complete a winning hand, and avoiding it pays only when an opponent is close to winning. If the in-house population creates those states at a different rate from the deployed population, a verdict on a defensive change is computed at the wrong exposure, whatever the sample size. Figure 1 shows the gap for one agent and what it does to an in-house verdict.

We measure this *population-exposure mismatch* in a Chinese Standard Mahjong (MCR) agent and ask what it implies for a concrete defensive intervention. Our contributions are:

- **An exposure measurement with explicit denominators** (Section 3). For one fixed focal policy, the per-game deal-in incidence is 16.6% in fully logged in-house self-play on a

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.

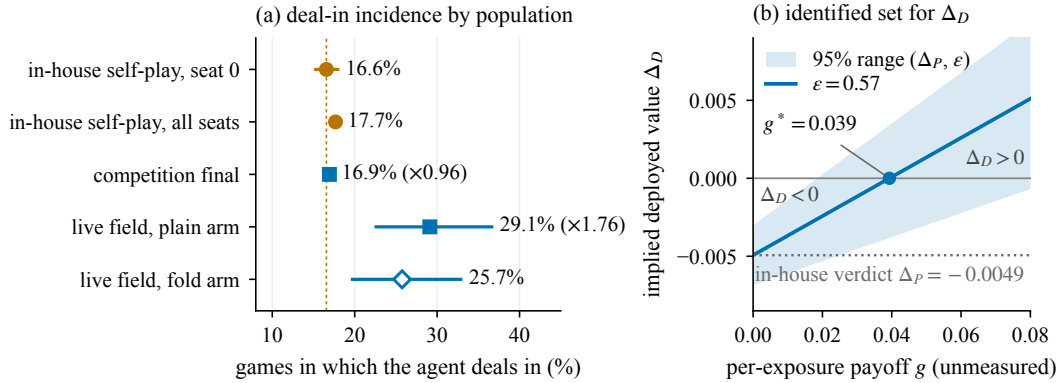

Figure 1: **The exposure gap and what it leaves identified.** **(a)** Per-game deal-in incidence of the deployed policy in three populations (the fold arm adds the fold overlay), with 95% Wilson intervals where games are independent; the dashed line marks the in-house seat-0 rate, and × gives the ratio to the in-house rate (seat 0 for the live field, all seats for the final). In-house rates come from our simulator with its scoring defect corrected (Section 2.1). The final carries no interval because its games share walls across seat permutations. **(b)** The deployed value implied by Equation 2 as a function of the unmeasured per-exposure payoff g, with D the live plain arm; the band propagates the 95% intervals of Δ_P and ε. The line crosses zero at $g^* = 0.039$. Without a bound on g, both signs remain possible.

corrected simulator, 16.9% in a competition final and 29.1% in a live-field arm. We split each rate into opponent completion and conditional deal-in, and keep this event rate distinct from the threat-state occupancy of the formal argument.

- **An identification analysis with stated assumptions** (Section 2.2). Under a population-invariant per-exposure payoff and cost (Assumptions 1–2), the deployed value of an intervention equals the in-house verdict plus an unobserved term proportional to the exposure gap. Against the live field ($\varepsilon = 0.57$, 95% CI $[0.44, 0.74]$) the identified set contains both signs and the payoff at which the sign flips is small; against the final $\varepsilon \approx 1$ and the in-house sign carries over.

- **A defensive intervention under three instruments** (Section 4). A fold policy is CI-separated negative in a calibrated in-house gate, run before the simulator fix, while changing 1.8% of discard decisions; a live A/B moves deal-ins in the expected direction without resolving it, and a power analysis bounds what a field A/B can resolve. Across seven overlay settings, the in-house verdict falls as engagement rises.

- **A paired-wall and event-level reading of a competition final** (Section 5). The duplicate format splits a statistically tied top pair into additive score channels, and a reconstruction of every discard from the game logs shows the finalists equally exposed to threat states but differing in conditional risk.

Related evaluation problems. Ratings and their robust alternatives (Elo, 1978; Herbrich et al., 2006; Balduzzi et al., 2018; Omidshafiei et al., 2019) concern how outcomes aggregate into rankings; exposure mismatch concerns which states the evaluation population generates, and a perfectly rated population can still have $\varepsilon \ll 1$ on a channel. Adversarial policies (Gleave et al., 2020; Wang et al., 2023) reach states that self-play rarely visits, from the attacker's side. ZSC-Eval (Wang et al., 2024), GAMMA (Liang et al., 2024) and risk-diverse self-play (Jiang et al., 2024) construct more diverse partner or opponent populations; the exposure statistic is a per-channel check of whether such a population covers the states where one skill pays. Duplicate formats and control variates (Zinkevich et al., 2006; White & Bowling, 2009; Burch et al., 2018) make evaluation efficient but do not choose the sampled population, and Mahjong agents (Li et al., 2020; Gao et al., 2019; Li et al., 2024; Lu et al., 2023) report validation accuracy, ladder or ranked play; we know of no prior measurement of how often in-house and deployed Mahjong populations punish one policy.

Table 1: Terms used throughout.

term	meaning
deal-in	the focal agent discards the tile that completes an opponent's winning hand
threat state S_C	a state in which an opponent is close enough to winning that a discard can deal in
exposure $\rho_P(S_C)$	occupancy of threat states under evaluation population P
deal-in incidence	fraction of games with a deal-in; depends on ρ_P *and* on the policy's conditional risk
ε	exposure ratio $\rho_P(S_C)/\rho_D(S_C)$, in-house population P to deployed field D
Δ_P, Δ_D	candidate-minus-reference placement difference under P and under D
g, c_P	payoff per unit of exposure, and expected cost (false folds, lost wins) under P

What we do not claim. We do not claim that self-play or behavioural cloning is unable in general to measure defensive skill, that the fold policy is valuable, or that the in-house verdict has the wrong sign. The claim is narrower: in the studied system the in-house population's exposure is below the live field's, though not the final's, and at that difference the in-house sign does not license a conclusion about the deployed sign. The in-house gate is sound for what it measures and remains the most precise instrument we have for skills the in-house population exercises (Section 6.1).

2 Setting and what an in-house verdict identifies

2.1 Game, agent and instruments

Game and agent. MCR is a four-player imperfect-information tile game. Tournament play uses a *duplicate* format: a fixed wall (the shuffled tile sequence) is replayed across seat permutations and placement points (4/3/2/1) are awarded on summed scores, which cancels much of the deal luck. The evaluated agent is a residual convolutional policy over a $(38, 4, 9)$ state tensor and 235 actions, trained by behavioural cloning on released MCR game records and deployed on a public online platform. The deployed policy, a three-student distillation ensemble averaging the students' action probabilities, is the *focal policy* of every exposure measurement.

The calibrated duplicate gate. A candidate is rotated through every seat assignment against three copies of the reference on fixed seeded walls, and summed scores are converted to placement points. When the candidate *is* the reference the gate returns exactly 2.500, so any drift of this identity cell is an evaluation error rather than a finding. Games are grouped into blocks of 2,000 on disjoint seed ranges, and verdicts are across-block means with t-based 95% intervals, of half-width about 0.002 placement points in the fold study below (12 blocks, $24{,}000$ games per arm). Placement points are 4 for first, so higher is better.

A scoring defect in the in-house simulator, and its correction. Our duplicate simulator's fan counter removed the winning tile from the claimant's hand before requiring thirteen tiles. For a win on a discard the hand already holds thirteen tiles without the discard, so when the claimant also held a copy of the winning tile the check failed and the win was not scored: play continued as if no win had been declared. Self-drawn wins were unaffected. The one-line fix removes the winning tile only for self-drawn wins. We re-ran the exposure instrument of Section 3 with the fix, with the same policy and the same $3{,}000$ seeds. An unpatched control re-run reproduces the original run count for count, so the fix is the only change: under the defect seat 0 dealt in in 14.07% of games (422), with the fix in 16.57% (497). The fix changes the outcome of 584 games, each now ending on a scored win on a discard (282 had ended on a different one, 234 on a self-draw and 68 in a draw). All in-house rates below come from the corrected run. The in-house verdicts of Section 4 (fold, divert and danger-penalty gates, and the danger-penalty incidences) were measured under the defect and were not re-run; only exposure was re-measured, and the defect's effect on those verdicts is unmeasured.

The deployed fields. We use two: the final of a recent international MCR AI competition, in which the evaluated agent was one of four finalists (512 walls $\times$ 24 seat permutations $=$ $12{,}288$ games, fully logged), and live play in the platform's public matchmaking against independently authored agents. Neither contains in-house agents other than the evaluated one.

2.2 EXPOSURE AND WHAT AN IN-HOUSE VERDICT IDENTIFIES

Let a candidate π' differ from a reference π on a *skill channel*: it acts differently on threat states S_C, and pays a cost for doing so elsewhere (a fold policy abandons some hands that would have won). An evaluation population P induces a state distribution ρ_P, and the value of the change under P is

$$\Delta_P \;=\; \rho_P(S_C)\, g_P \;-\; c_P, \tag{1}$$

where g_P is the expected payoff of the changed behaviour per unit of threat exposure and c_P is its expected cost. Equation 1 is an accounting identity once g_P and c_P are defined this way. Relating two populations requires assumptions:

Assumption 1 (Payoff invariance)**.** *$g_P = g_D = g$: a threat state, once reached, pays the same on average under P and D.*

Assumption 2 (Cost invariance)**.** *$c_P \approx c_D$: the intervention's engagement rate and the cost per engagement (false folds, lost wins) are similar under P and D.*

Assumption 3 (Incidence as an exposure proxy)**.** *For a fixed focal policy, the probability of dealing in given a threat state is the same under P and D, so the ratio of deal-in incidences estimates $\varepsilon = \rho_P(S_C)/\rho_D(S_C)$.*

We do not measure engagement, false folds or lost wins in the deployed field, so Assumption 2 is an assumption, not a finding. Assumption 3 is the weakest link: the conditional deal-in rate given an opponent's completed hand is close but not equal across populations (Section 3).

Proposition 1 (Identified set)**.** *Under Assumptions 1 and 2, with $\rho_P(S_C) = \varepsilon\rho_D(S_C)$,*

$$\Delta_D \;=\; \Delta_P + (1-\varepsilon)\, g\, \rho_D(S_C). \tag{2}$$

If $\varepsilon < 1$ and $\Delta_P < 0$, then $\Delta_D > 0$ if and only if $g > g^* = |\Delta_P|/\big((1-\varepsilon)\rho_D(S_C)\big)$. With $g \geq 0$ otherwise unconstrained, the identified set for Δ_D is $[\Delta_P, \infty)$, which contains both signs. If $\varepsilon \geq 1$, then $\Delta_D \leq \Delta_P$ for every $g \geq 0$.

The proof is one line: subtract Equation 1 under P from the same identity under D and use the assumptions. A negative in-house verdict on a low-exposure channel is evidence against the skill only together with an upper bound on g below g^*. In partial-identification terms (Manski, 1990; 2003; Tamer, 2010), without such a bound we can report when the in-house sign fails to license the deployed sign, not estimate the deployed value. In transportability terms (Pearl & Bareinboim, 2011; Bareinboim & Pearl, 2016), Equation 2 is the correction the in-house population cannot supply, because the shift is concentrated on the states where the two policies differ. More games shrink the interval around Δ_P, not around Δ_D, and control variates (White & Bowling, 2009; Burch et al., 2018) reduce $\mathrm{Var}(\hat{\Delta}_P)$ without changing what the gate estimates.

Worked example. The fold study of Section 4 gives $\Delta_P = -0.0049$, measured under the simulator defect. Taking the live plain arm as D (ρ_D proxied by its 29.1% incidence) and corrected in-house self-play as P (16.6%, from a fully logged re-run of the gate's plain-arm pairing), $\varepsilon = 0.166/0.291 = 0.57$ and $g^* = 0.00493/(0.431 \times 0.291) = 0.039$ placement points per unit of exposure. Propagating the 95% intervals of ε and of Δ_P separately moves g^* within $[0.030, 0.064]$ and $[0.024, 0.055]$. The fold policy's deployed value is positive only if an exposed game gains it more than g^*, about four hundredths of a placement point on a scale spanning three points. Figure 2 draws the identity with these numbers: all payoffs agree at the in-house exposure, which is why the gate cannot tell them apart, and fan out by the live field's; no measurement of g excludes values on either side of the crossing. Under every choice of P and D (Figure 3), g^* is small when the gap to D is large and grows without bound as $\varepsilon \to 1$, and at $\varepsilon \geq 1$ a negative Δ_P implies a negative Δ_D.

3 EXPOSURE: IN-HOUSE MATCHES THE FINAL, NOT THE LIVE FIELD

In-house population. We re-ran the strongest in-house field we can construct, the deployed ensemble in all four seats, for 3,000 duplicate-seeded games on the corrected simulator, with every outcome and discarder logged. The population is not passive: 97.7% of games end in a completed hand (mean 13.0 fan). Seat 0 deals in in 16.6% of games (95% CI $[15.3, 17.9]$) $= 69.1\% \times 23.97\%$; pooling all seats gives $17.7\% = 73.3\% \times 24.11\%$ (95% CI $[17.3, 18.1]$ over games, as a game has at most one deal-in). Seats are not exchangeable within a deal, so we report both views.

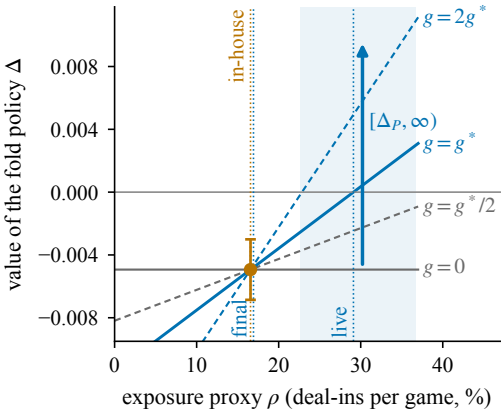

Figure 2: **Proposition 1 as a worked example.** Every payoff g gives a line through the in-house verdict (orange, 95% interval) at the in-house exposure; its value at the live field's exposure (shaded: 95% interval) is the deployed value. The line with $g = g^*$ crosses zero there; the arrow is the identified set.

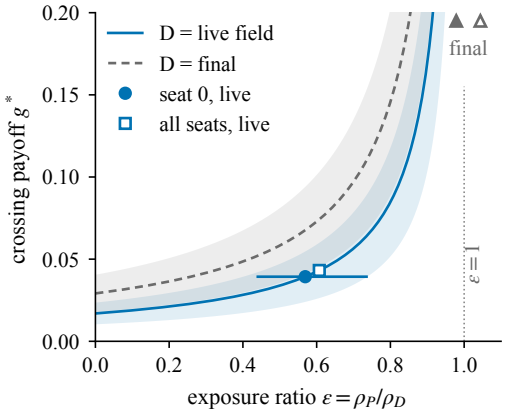

Figure 3: **Sensitivity of ε and g^* to the choice of populations**, with $\Delta_P = -0.0049$ (bands: its 95% interval). Markers: in-house view (seat 0, all seats) against each deployed field; bar: 95% interval of ε. Against the final $\varepsilon \approx 1$ (triangles, filled for seat 0): the seat-0 crossing is off scale ($g^* = 1.4$); with all seats $\varepsilon > 1$ and there is none.

Table 2: **Deal-in incidence with explicit denominators**, for the deployed ensemble (the fold arm adds the fold overlay). Per-game incidence factors as opponent completion (fraction of games an opponent completes a hand) times conditional deal-in (fraction of those hands completed on the focal agent's discard). "All seats" pools the four identical-policy seats of each self-play game. The live-field rows are the two arms of the A/B of Section 4; completion counts were not logged there. In-house rows come from the corrected simulator (Section 2.1).

population (focal: deployed ensemble)	games	opp. completion	cond. deal-in	deal-in / game
in-house self-play, seat 0	3,000	69.1%	23.97%	16.6%
in-house self-play, all seats	3,000	73.3%	24.11%	17.7%
competition final (seat-rotated)	12,288	73.4%	23.08%	16.9%
live field, plain arm	158	–	–	29.1%
live field, fold arm	167	–	–	25.7%

Competition final. Of the 12,288 games, 98.4% ended in a completed hand and the evaluated agent won 25.0%, so an opponent completed the hand in 73.4%; at 16.9% per-game incidence, the evaluated agent was the discarder in 23.08% of opponent-completed hands. The final's incidence is 0.96 times the all-seat in-house rate (1.00 in opponent completion, 0.96 in conditional deal-in) and 1.02 times the seat-0 rate, inside its interval: in-house self-play and the final expose the deployed policy equally often.

Live field. In the live A/B the plain arm, which runs the same deployed ensemble, deals in in 29.1% of games (95% CI $[22.6, 36.6]$, n=158), 1.76 times the in-house seat-0 rate (95% CI $[1.36, 2.27]$). This gives the worked example's $\varepsilon = 0.57$ (95% CI $[0.44, 0.74]$ by the delta method on the log ratio; 0.61 with the all-seat rate). Opponent completion was not logged live, so the two factors cannot be separated there. Against the final, $\varepsilon = 0.98$ (seat 0) or 1.04 (all seats), so Equation 2 crosses zero at $g^* = 1.4$ or not at all: relative to the final, the negative in-house verdict transports unless an exposed game gains more than 1.4 placement points. The mismatch, and with it the identification problem, is a property of the pair of populations, not of self-play as such (Figure 3).

Table 3: **The fold policy under three instruments.** Placement is fold minus plain; brackets are 95% intervals. The engagement counters come from the gate's own games; the live A/B draws opponents from the platform's matchmaking; the gate ran before the simulator fix (Section 2.1). [a]The gate's deal-in counter did not record scored deal-ins (Section 4). [b]The live records do not state whether higher placement is better.

instrument	games per arm	deal-ins per game	placement, fold − plain
calibrated gate (in-house)	24,000	not recorded[a]	-0.0049 $[-0.0069, -0.0030]$
engagement counters	same	1.8% of discards changed	–
live A/B	158 / 167	$29.1\% \to 25.7\%$	-0.022 $[-0.25, +0.21]$[b]

4 A DEFENSIVE INTERVENTION UNDER THREE INSTRUMENTS

The intervention is a rule-based fold overlay on the deployed ensemble. After its tenth discard of a deal, the overlay latches into a folded state for the rest of the deal when its best hand is at least two tiles from ready while an opponent has declared two or more melds, or at least three tiles from ready regardless. While folded it discards the safest legal tile by a fixed ordering (tiles the threatening opponents have already discarded, then honours with three copies visible, then tiles with more copies visible, honours and terminals before middle tiles) and passes on every claim except a win. Table 3 summarises three instruments.

4.1 THREE INSTRUMENTS

Instrument 1: the calibrated gate. In-house, paired against the plain ensemble, the fold policy gates at -0.0049 placement points (SE 0.0009 across 12 blocks; $t = -5.6$, $p = 1.5 \times 10^{-4}$). The verdict is negative in 11 of 12 blocks (sign test $p = 0.006$). The interval treats blocks, drawn from disjoint seed ranges, as independent (per-block values in Appendix A).

Instrument 2: engagement counters. In the same games the overlay changes 1.8% of the candidate's discard decisions ($5{,}525$ of $301{,}417$), latches $1{,}659$ times and passes on 376 claims, so the negative verdict is not a no-op. The gate's deal-in counter cannot be read as deal-ins: it was incremented only when an opponent's policy, queried about a claim on the candidate's discard, declared a win, whereas the simulator scores a legal win on a discard before any query. Its 906 and 871 events (plain, fold) are therefore win declarations the simulator did not score, the defect of Section 2.1 seen from the policy's side; exposure comes from the corrected re-run of Section 3 instead.

Instrument 3: a live A/B. We ran fold against plain in the live field under a fixed protocol (325 completed games: 167 fold, 158 plain, no errors or unfinished games); later games of this A/B survive only as a prose summary and are not used. The live arm used the deployment build of the overlay, a separate implementation of the same rule family whose settings were not stored with the A/B records. Deal-in incidence fell from $46/158 = 29.1\%$ to $43/167 = 25.7\%$, a 3.4-point (12% relative) difference with 95% CI $[-6.3, 13.1]$ points ($p = 0.50$): the direction the policy is designed for, and no more. Mean placement was 2.557 (fold) and 2.579 (plain), a difference of -0.022 with SE 0.116 (95% CI $[-0.25, +0.21]$). The live records do not state their placement convention, so which arm this favours is unknown; with an interval spanning zero, the field does not identify the sign of the placement effect under either convention.

How much field data would resolve it. With the A/B's per-game placement standard deviation of 1.05, 80% power at two-sided $\alpha = 0.05$ needs about $2{,}800$ games per arm for the observed deal-in difference, $6{,}900$ for a placement difference of 0.05, $43{,}000$ for 0.02 and $710{,}000$ for one the size of the in-house verdict (0.005). A feasible field A/B checks direction; pricing the channel in placement needs long-horizon ranked play or a direct estimate of g.

Reading the three together. The gate gives a precise negative verdict on a population whose exposure is 0.57 of the live field's (95% CI $[0.44, 0.74]$). Proposition 1 makes this compatible with a positive deployed value whenever g exceeds $g^* = 0.039$, and neither the counters nor the A/B bound g, so the deployed sign is *not identified*. This is weaker than a demonstrated false negative and stronger than "no effect detected": the in-house negative is real under the simulator's rules as they stood when it was measured, defect included, and alone says nothing about the live-field sign.

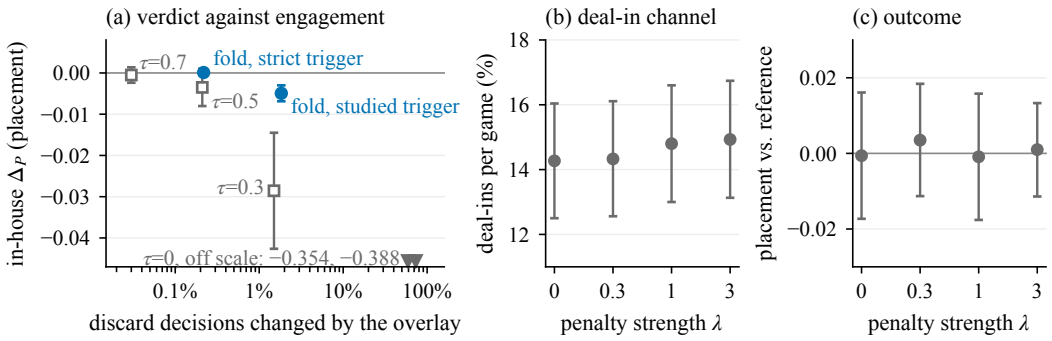

Figure 4: **In-house verdicts of defensive variants.** **(a)** Gate placement against the fraction of discard decisions changed: the fold overlay at the studied and a stricter trigger (blue, t intervals over 12 blocks) and a top-K divert overlay at five danger thresholds τ (grey, 4 blocks; the two $\tau = 0$ settings lie off scale). **(b, c)** A danger penalty of strength λ in distillation: deal-in incidence (1,500 games) and placement against a single-model reference, with 95% intervals; values in Appendix B. All values predate the simulator fix (Section 2.1).

4.2 Trigger, engagement and other defensive variants

Figure 4 collects every in-house verdict on a defensive change that we have. A stricter trigger for the same overlay (three and four tiles from ready instead of two and three) latches 230 times, changes 0.22% of discard decisions (8 times fewer) and ties ($+0.0001$, 95% CI $[-0.0005, +0.0007]$). A top-K divert overlay, in which a per-candidate danger model (recorded AUROC 0.97 on labels from the same simulator) diverts the chosen discard to a safer one among the policy's top K when its predicted danger exceeds τ, ties or loses at every threshold. Across the 7 settings the verdict falls as engagement rises (Spearman $\rho = -0.86$; -1.0 within the divert family), as Equation 1 predicts when the cost is paid on every changed decision and the benefit only on exposed states. It is equally consistent with overlays that are simply worse, and the families differ in level at similar engagement. A training-time danger penalty moved no in-house deal-ins at any strength (Figure 4b; the unpenalised control deals in in 14.3% of games under the defect), so its tied placements are not evidence about defence: the channel never moved.

5 Competition final: a tie that splits into channels

The final gives a large, fully logged comparison of four independently built agents on identical walls. We name them by their rank on total score and, to preserve anonymity, do not say which of them is the evaluated agent. On total score the top three are statistically indistinguishable: wall-level paired tests over the 512 walls give $|t| \leq 0.36$ for every top-three pair. The fourth agent trails each of them (t between 2.00 and 2.65, unadjusted $p = 0.008$–0.046), but Holm-adjusted over the six pairwise tests $p = 0.050$–0.19, so even that separation is marginal. The agent third on total score has the best mean placement; we do not test placement differences. No agent had an error or time-limit ending.

5.1 The paired decomposition

For the first and second agents, the 2,048 matched cells in which both played the same seat on the same wall give a paired score difference per cell. Writing a cell's score as win points plus the (negative) payments made, the per-cell difference is exactly additive:

$$\Delta\text{score} \;=\; \Delta\text{win points} \;+\; \Delta\text{deal-in payments} \;+\; \Delta\text{other payments}, \tag{3}$$

where "other payments" are payments on opponents' self-draws plus bystander payments. Averaged over cells (Figure 5c; first minus second, per game), the components are $+0.08$ in win points, -0.10 in deal-in payments and $+0.07$ in other payments: a total of $+0.05$ (paired SE 0.364, t=0.13), with 879 cells to the first, 880 to the second and 289 ties. The title margin is a tie that splits into opposing channels: the two agents win equally often (difference -0.001), the first converts more

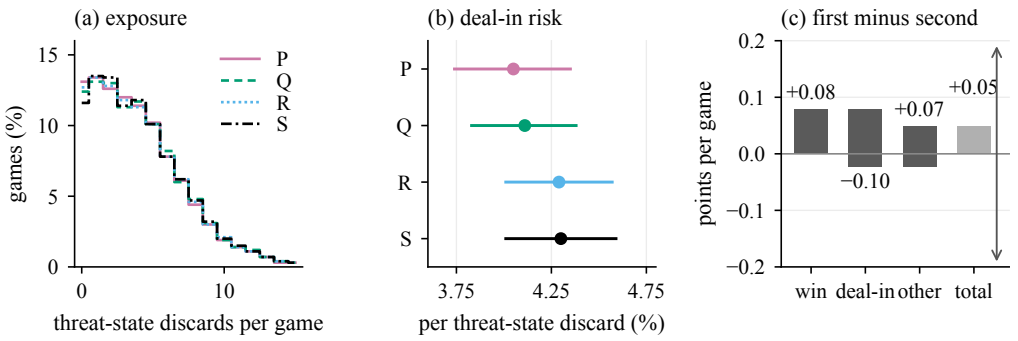

Figure 5: **The defensive channel in the final. (a)** Threat-state discards per game and **(b)** deal-ins per threat-state discard (95% intervals clustered by wall), per finalist over all 12,288 games; finalists are labelled P–S in increasing order of (b), unrelated to rank. **(c)** Paired decomposition, first minus second on total score; the total's 95% interval, $[-0.67, +0.76]$, extends beyond the axis.

of its wins into self-draws, and the second pays less for deal-ins. On the 835 cells in which both won, the second's winning hands are not smaller (12.61 vs 12.48 fan). The fourth agent's deficit relative to the first (-1.06 points per game; t=2.52) splits into -0.79 in win points, -0.08 in deal-in and -0.19 in other payments (every finalist played every game, so these unpaired means are balanced): the largest gap in the final is mostly offensive. The decomposition is accounting, not a counterfactual; the retained summary has no per-cell vectors, so the components carry no standard errors, each is smaller than the SE of the total, and we read their signs.

5.2 The deal-in channel, discard by discard

To see the deal-in channel at the level of single decisions, we reconstructed every hand in the final from the game logs and marked each discard made while some opponent was one tile from a complete hand (shanten 0 in the standard, seven-pairs or thirteen-orphans form, ignoring the 8-fan minimum). Every discard is found in the reconstructed hand, the winner's hand before the winning tile is one tile from complete in 99.8% of wins outside the knitted forms the indicator does not track (5.0% of wins), and 97–98% of each finalist's deal-ins come from a marked discard.

On this measure the four finalists are equally exposed, with 3.95–4.03 threat-state discards per game (Figure 5a), and differ in the risk per such discard, 4.05–4.30% (Figure 5b). Within one field the exposure is shared and the defensive channel acts through conditional risk; Assumption 3 asks the reverse across populations. For the top pair the channel can also be read on identical deals. In each of the 6,144 swap pairs the two agents exchange seats on the same wall, with the same agents in the other two seats; read from both exchanged seats, the first alone deals in on 1,016 occasions and the second alone on 961 (wall-level t=0.86, p=0.39), so the deal-in component of the tie is not separable decision by decision either.

6 Discussion and Conclusion

6.1 An evaluation recipe, split by what it protects

We separate practices that address exposure validity, the subject of this paper, from general hygiene that any gate needs. They are diagnostic: they say when a verdict should not be trusted, not how to correct it, and we have not shown that following them selects better agents.

Exposure validity. (E1) Report the exposure with the verdict. For each channel a candidate is meant to move, measure its exposure in-house and in the field with explicit denominators (Table 2), and report ε and g^* beside the verdict. By Proposition 1, a negative in-house verdict at $\varepsilon < 1$ is evidence against the skill only together with a bound on g. **(E2) Price low-exposure channels where the exposure is, with a power calculation first.** Feasible field A/Bs check direction (Section 4); placement value needs thousands of games per arm or long ranked play. **(E3) Decompose paired**

comparisons into channels and events. On shared walls, Equation 3 says which channels carried a verdict, and reconstructed threat states separate exposure from conditional risk (Section 5).

(E4, proposed and not evaluated here) Stratify the gate by exposure. If each in-house game is labelled by whether a threat state occurred, the gate can report paired differences on exposed and unexposed games, $\hat{\Delta}_P^{\mathrm{exp}}$ and $\hat{\Delta}_P^{\mathrm{unexp}}$, and re-weight them to the field's exposure, $\hat{\Delta}_D = \hat{\rho}_D\,\hat{\Delta}_P^{\mathrm{exp}} + (1-\hat{\rho}_D)\,\hat{\Delta}_P^{\mathrm{unexp}}$. This post-stratification replaces Assumptions 1–2 by the weaker requirement that the *stratum-conditional* effects transport, and needs only a field estimate of exposure, which the event reconstruction of Section 5 supplies from game logs. Its variance grows as the exposed stratum shrinks in-house: the mismatch reappears as a precision cost rather than a bias. Our gate records carry no per-game exposure labels, so it is the first experiment we would run.

General hygiene. (H1) Calibrate the identity cell. Candidate = reference must return the theoretical tie exactly; analyse in blocks and report intervals across blocks. **(H2) Report engagement counters, and check counters and simulator against the events they name.** A null with no engagement, or with no movement of the channel, is not a negative result about the skill (Figure 4b). A counter must be validated against the scored events, and a simulator's scoring against the official rules event by event; neither check preceded our gate verdicts (Section 2.1). For decisions the in-house population exercises at field-like rates, such as discard efficiency, the gate remains the most precise instrument available (half-width about 0.002 placement points here): gate such channels in-house, count mechanisms everywhere, and price low-exposure channels in the field.

6.2 Limitations and scope

All evidence is from one game, one platform and one agent lineage; the identification argument is general, but the exposure values and the fold verdicts are facts about this system. Deal-in incidence is a proxy for threat-state occupancy, and Assumption 3 is only partly supported: conditional deal-in in the final is 0.96 times the in-house rate, it was not logged live, and event-level threat states exist for the final only. We did not measure engagement, false folds or lost wins in the field, so cost invariance (Assumption 2) is assumed. The exposure difference is not causally isolated from other population differences; the missing experiment is a dose-response study that holds the candidate fixed, varies the population across the exposure range (for example with risk-diverse populations) and tests whether the in-house verdict tracks ε. In-house games run on our simulator and field games on the official judges. We corrected the one scoring difference we found (Section 2.1) for the exposure measurement, but every in-house gate verdict predates the fix and was not re-run, and the simulator has not been checked against the official judges. The live overlay is a separate implementation from the gated one, the live A/B is small, the decomposition lacks component uncertainty, and E4 is untested. Beyond this game the argument needs a skill whose payoff is concentrated on states that opponents create, a counter for them and a paired gate; defence against rare, costly lines of play has this shape elsewhere, but we have measured exposure in no other domain.

6.3 Conclusion

In this work, we measured how often in-house and deployed populations expose one Mahjong policy to the states in which defence pays, and asked what an in-house verdict on a defensive change identifies there. With a scoring defect in our simulator corrected, the in-house rate is 0.57 of the live field's and matches a competition final's. Under stated assumptions for transporting a verdict, a precise negative in-house verdict leaves the deployed sign unidentified at any exposure below the field's, as against the live field, and carries over where exposure matches, as against the final; a live A/B and a power analysis bound what feasible field data can add. In the final, a paired decomposition and an event-level reconstruction show the defensive channel at work among strong agents: equal exposure, different conditional risk. Reporting the exposure ratio and the crossing payoff next to a verdict costs one counter per channel and says when not to trust it, once counter and simulator are checked against the official rules; we found our simulator's defect only after the gate verdicts, which were not re-run. We hope that exposure-stratified gates and dose-response studies with risk-diverse populations will make such verdicts correctable, not only diagnosable.

REPRODUCIBILITY STATEMENT

The supplementary material contains the result files behind every number in the paper (exposure counts and the per-game rows of the corrected re-run and of its unpatched control, the one-line simulator fix, gate blocks for both fold triggers with their engagement counters, the divert and penalty gates, live A/B aggregates, competition results, the paired decomposition, and the threat-state events derived from the final's game logs), the script that derives every computed statistic in the text from those files (intervals, ratios, crossings, paired tests, clustered rates and power calculations), the figure scripts, the instrument used for the exposure re-run and the script that reconstructs threat states from game logs. For anonymity the competition final is released as a reduced record: finalists appear by rank on total score in pairwise quantities or under labels unrelated to rank in per-finalist rates, the evaluated agent is not marked, and match identifiers, wall seeds and per-wall records are removed. The game engine, model checkpoints and raw game logs are not included, so the games cannot be re-played from the supplement; every printed number is stored in the released files or recomputed from them.

AI USE STATEMENT

We used generative AI tools on the following tasks from ICLR's required-disclosure list: implementing analysis code, helping to interpret results, and giving feedback on the framing of claims and their assumptions. We did *not* use generative AI tools to generate data reported as measurements; every game, gate run and field match reported here was played by the agents described. From the recommended-disclosure list we additionally used generative AI tools to draft and edit parts of this paper, to create figures from the result files, to identify and summarise related literature, to check references against their primary records, and to format references.

We have reviewed all AI-assisted work. Every number in the paper was recomputed by a script from the released result files and compared against them; every added citation was checked against its arXiv or publisher record; claims that could not be traced to a result file were removed. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

REFERENCES

David Balduzzi, Karl Tuyls, Julien Pérolat, and Thore Graepel. Re-evaluating evaluation. In *Advances in Neural Information Processing Systems 31*, pp. 3268–3279, 2018.

Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, Iain Dunning, Shibl Mourad, Hugo Larochelle, Marc G. Bellemare, and Michael Bowling. The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence*, 280:103216, 2020.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Neil Burch, Martin Schmid, Matej Moravčík, Dustin Morill, and Michael Bowling. AIVAT: A new variance reduction technique for agent evaluation in imperfect information games. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 949–956, 2018. doi: 10.1609/aaai.v32i1.11481.

Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-AI coordination. In *Advances in Neural Information Processing Systems 32*, pp. 5175–5186, 2019.

Wojciech M. Czarnecki, Gauthier Gidel, Brendan Tracey, Karl Tuyls, Shayegan Omidshafiei, David Balduzzi, and Max Jaderberg. Real world games look like spinning tops. In *Advances in Neural Information Processing Systems 33*, pp. 17443–17454, 2020.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Publishing, New York, 1978.

Shiqi Gao, Fuminori Okuya, Yoshihiro Kawahara, and Yoshimasa Tsuruoka. Building a computer mahjong player via deep convolutional neural networks. *arXiv preprint arXiv:1906.02146*, 2019.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *International Conference on Learning Representations*, 2020.

Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems 19*, pp. 569–576, 2006.

Yuhua Jiang, Qihan Liu, Xiaoteng Ma, Chenghao Li, Yiqin Yang, Jun Yang, Bin Liang, and Qianchuan Zhao. Learning diverse risk preferences in population-based self-play. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, 2024.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems 30*, pp. 4190–4203, 2017.

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

Xiali Li, Bo Liu, Zhi Wei, Zhaoqi Wang, and Licheng Wu. Tjong: A transformer-based mahjong AI via hierarchical decision-making and fan backward. *CAAI Transactions on Intelligence Technology*, 9(4):982–995, 2024. doi: 10.1049/cit2.12298.

Yancheng Liang, Daphne Chen, Abhishek Gupta, Simon S. Du, and Natasha Jaques. Learning to cooperate with humans using generative agents. In *Advances in Neural Information Processing Systems 37*, 2024.

Yunlong Lu, Wenxin Li, and Wenlong Li. Official international mahjong: A new playground for AI research. *Algorithms*, 16(5):235, 2023.

Charles F. Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2): 319–323, 1990.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, New York, 2003.

Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M. Czarnecki, Marc Lanctot, Julien Pérolat, and Rémi Munos. α-rank: Multi-agent evaluation by evolution. *Scientific Reports*, 9:9937, 2019.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 247–254. AAAI Press, 2011.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

Elie Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2:167–195, 2010.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354, 2019.

Tony T. Wang, Adam Gleave, Tom Tseng, Kellin Pelrine, Nora Belrose, Joseph Miller, Michael D. Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, and Stuart Russell. Adversarial policies beat superhuman Go AIs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023.

Xihuai Wang, Shao Zhang, Wenhao Zhang, Wentao Dong, Jingxiao Chen, Ying Wen, and Weinan Zhang. ZSC-Eval: An evaluation toolkit and benchmark for multi-agent zero-shot coordination. In *Advances in Neural Information Processing Systems 37 (Datasets and Benchmarks Track)*, 2024.

Martha White and Michael Bowling. Learning a value analysis tool for agent evaluation. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence*, pp. 1976–1981, 2009.

Martin Zinkevich, Michael Bowling, Nolan Bard, Morgan Kan, and Darse Billings. Optimal unbiased estimators for evaluating agent performance. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pp. 573–579, 2006.

A PER-BLOCK GATE VALUES FOR THE FOLD POLICY

Paired placement differences (fold minus plain) in the 12 blocks of 2,000 games per arm, in block order, at the studied trigger (all gate values predate the simulator fix of Section 2.1): -0.0050, -0.0075, -0.0115, -0.0038, -0.0053, $+0.0002$, -0.0035, -0.0055, -0.0055, -0.0053, -0.0060, -0.0005. Mean -0.0049, SE 0.0009, 95% t interval (11 df) $[-0.0069, -0.0030]$. At the stricter trigger, on the same walls: -0.0015, -0.0008, -0.0002, $+0.0008$, $+0.0013$, $+0.0002$, 0.0000, $+0.0005$, -0.0013, $+0.0013$, $+0.0002$, $+0.0008$; mean $+0.0001$, SE 0.0003, t=0.4, negative in 4 of 12 blocks. Block start seeds are $1{,}000$ apart, and each block uses 500 walls in four seat rotations, so the seed ranges are disjoint by construction; the retained records do not include per-wall lists, so this is the design rather than an audit.

B VALUES OF THE DEFENSIVE VARIANTS

Table 4: **Values plotted in Figure 4.** Left: top-K divert overlay, absolute gate placement against the plain ensemble (tie $= 2.500$; four blocks of 2,000 games on disjoint seeds). Right: two-student ensembles distilled with a danger penalty of strength λ, gated against a single-model reference (8 blocks, $16{,}000$ games) with deal-in incidence from $1{,}500$ games; brackets are 95% CIs. All values predate the simulator fix (Section 2.1).

divert setting	overrides	placement
$\tau = 0.7$, top 3	0.03%	2.4995 [2.4976, 2.5014]
$\tau = 0.5$, top 3	0.21%	2.4965 [2.4920, 2.5010]
$\tau = 0.3$, top 3	1.5%	2.4714 [2.4574, 2.4855]
$\tau = 0$, top 3	59.2%	2.1458 [2.1302, 2.1613]
$\tau = 0$, all legal	73.2%	2.1117 [2.1014, 2.1220]

λ	placement	deal-in / game
0	2.4994 [2.4827, 2.5161]	14.27% [12.50, 16.04]
0.3	2.5035 [2.4887, 2.5184]	14.33% [12.56, 16.11]
1.0	2.4991 [2.4824, 2.5158]	14.80% [13.00, 16.60]
3.0	2.5010 [2.4886, 2.5133]	14.93% [13.13, 16.74]

C STATISTICAL METHODS

Binomial intervals are Wilson intervals. The live deal-in difference uses a normal interval for a difference of independent proportions; deal-in counts are recovered from the stored rates and arm sizes ($43/167$ and $46/158$ reproduce the stored rates to three decimals). The all-seat in-house interval is a

Wilson interval on the games with a deal-in (at most one per game), divided by four. The interval for ε applies the delta method to the log ratio of two independent proportions (live plain arm; in-house seat 0). Fold-gate intervals are t intervals over blocks. The six wall-level pairwise tests of the final are adjusted with Holm's step-down procedure. Power calculations use the normal approximation for two independent arms with the A/B's pooled per-game standard deviation (from the stored standard errors and arm sizes) and, for deal-ins, the observed arm rates. Competition-final paired tests are over the 512 walls, each summing its 24 seat permutations; per-game values divide by 24. The threat-state rates of Figure 5a,b are ratios of sums with delta-method intervals clustered by wall (512 clusters), and the swap-pair comparison is a paired t test over walls of the per-wall difference in discordant pairs. The Spearman coefficient of Section 4 is descriptive (7 settings from two overlay families). All values are produced by one script from the released result files.