

# May the Best Bot Win

## *Variance and Sample Efficiency Across Poker Formats*

Henry De Courcy Thompson, FullHouse Research

### ABSTRACT

The length of simulations required to distinguish poker agents depends strongly on the format being played. We benchmark matched agent evaluation across seven poker environments and ask both how large the simulation burden becomes and how much of it can be removed through variance reduction. At a common operational target of 1 BB/100 and 80% power, median played-hand requirements range from roughly 0.34 million for 4-max All-in-or-Fold to 20.5 million for NLHE Bomb Pot, with 4-max and 6-max NLHE requiring about 4.3 to 4.4 million. These are conditional benchmark requirements, not intrinsic difficulty rankings of the games.

The cost differences are associated with highly concentrated matched payoff distributions. Across the 100BB environments, blocks containing a pot of at least one starting stack account for 79 to 97% of squared deviation. Controlled NLHE interventions show that simulation burden also depends on stack depth, table size and the chosen effect-size scale. At a fixed 1 BB/100 target, raw matched variability rises with stack depth, while a 3-to-8-seat sweep increases median played-hand requirements from 3.71 million to 5.47 million.

Variance reduction can materially change the cost, but not uniformly. Exact all-in EV removes about 90% of variance in 8BB All-in-or-Fold, 28% at 20BB NLHE and 12% at 100BB NLHE. At 100BB, fitted AIVAT removes about 42%, but its approximately 1.72 times sample-efficiency gain is offset by an approximately 1.86 times simulation slowdown in our implementation. Fixed-budget inference remains well calibrated at the practical sample sizes studied, while naive repeated monitoring does not. The results provide an empirical basis for choosing poker formats, simulation budgets and variance-reduction methods when reliable agent rankings must be produced under finite compute.

# Introduction

Every poker player has watched a fish run good, and every player has watched skill prevail over time, hence the old adage: “Nobody runs better than the pro”. When faced with the problem of ranking players, or poker bots, this raises a vital question: how long must we run a simulation to be confident enough to declare one participant the strongest? And how does this vary across different configurations, game modes and table sizes? Especially when large prizes, tournaments or scientific frontiers are at stake, we must be reasonably certain of our ranking, without spending unnecessary time and compute once the precision we need has been reached.

In the previous FullHouse Research note, we considered the awkward question of ranking criteria. However, even with a perfect ranking mechanism we need to know how long to run it before acceptance. Poker is incredibly noisy, skewed by lucky rivers, big pots, and variance at every street. A short sample can be extremely misleading.

Computer poker has dealt with the obvious version of this problem for years. The Annual Computer Poker Competition used duplicate poker, replaying the same underlying deals while changing which agent occupied each seat, so that comparisons depended less on the luck of the deal (Bard et al., 2013). In simulation language this is common random numbers: expose competing systems to the same random environment, then analyse their difference rather than two independent totals.

That removes a large amount of variance, but it does not make the difference between two agents stable. Equivalent opportunities can still produce very different strategic paths, and a small number of consequential pots can dominate the final mean. The size and shape of that residual distribution also depend on the environment used to generate it. Through our paired comparison environment, we attempt to answer a narrower question than cross-tournament rankings:

**Where does matched poker-agent evaluation noise come from, and how does it change with game structure, stack depth and table size?**

## Experimental framework

For each evaluation block, candidates A and B face the same underlying chance realization, equivalent seat positions and the same reference-agent configuration. Let  $X_{A,i}$  and  $X_{B,i}$  denote their returns from block  $i$ . The statistical observation is the paired difference

$$D_i = X_{A,i} - X_{B,i}.$$

The individual hands inside a block are deliberately coupled and are not treated as independent samples. Standard 4-max blocks contain about eight played hands and 6-max blocks about twelve; the table-size experiment generalises this construction to 3, 4, 6 and 8 seats. Inference is performed in blocks, while played hands are used only as an operational cost.

Our core benchmark suite contains 4-max All-in-or-Fold, 4-max and 6-max NLHE, Crazy Pineapple, PLO, NLHE Bomb Pot and the Nit Game. We supplement this suite with two controlled NLHE interventions: a 20/50/100BB stack-depth sweep and a 3/4/6/8-max table-size sweep at 100BB. Three non-control agent

pairings are used per environment, together with exact self-play controls.

That population choice is part of the experiment. We do not interpret a measured standard deviation as "the variance of PLO" or "the variance of NLHE". It is the variance induced by the game rules, stack structure, candidate pair and reference population together. Absolute hand requirements are therefore conditional on those populations and on how the target effect is defined.

For a concrete operational ruler we report the number of matched blocks, and the corresponding played hands, required to resolve a 1 BB/100 edge at a two-sided 5% test with 80% power. We use 1 BB/100 because it is familiar and interpretable in poker, not because it is a standardised effect size across variants or stack depths. The same empirical residual distributions are also resampled directly. At the analytically predicted 80% target, empirical power lies between 0.779 and 0.815 across the final pairings, so the large cross-environment differences are mainly differences in the scale of the matched distribution, not a catastrophic failure of fixed-N large-sample inference.

## Sources of variance

The residual variance is not spread evenly across matched blocks. Across the 100BB environments and agent pairings, blocks containing a pot of at least one starting stack account for 79 to 97% of total squared deviation. Many large-pot blocks still produce a zero matched difference, so the contribution is concentrated further in the subset where the two agents actually diverge.

The same concentration appears without choosing a pot threshold. Across the benchmark medians in Figure 1, the top 1% of matched blocks account for roughly 19 to 82% of squared deviation, with some individual pairings more concentrated still. A small fraction of blocks can therefore carry a highly disproportionate share of the uncertainty in the estimated mean.

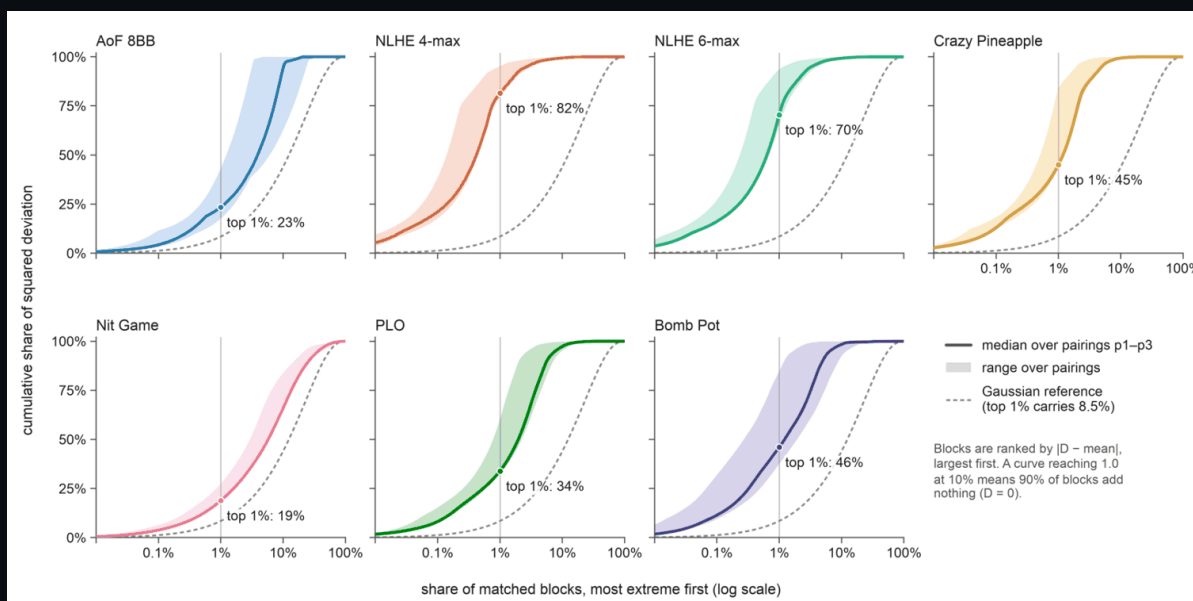


**Figure 1.** A small fraction of matched blocks can carry a disproportionate share of evaluation uncertainty. Each panel shows the cumulative share of squared deviation contributed by the most extreme blocks, with a Gaussian reference for comparison. Across the benchmark medians, the top 1% of blocks accounts for roughly 19 to 82% of squared deviation; the shaded ranges show variation across agent pairings. The table-size intervention later shows that this tail concentration need not increase when the number of seats increases.

All-in-containing blocks are a useful proxy for this mechanism, but not its definition. In PLO, 6.4% of blocks contain a pot of at least one starting stack without an all-in, and those blocks carry a further 12 to 13% of squared deviation. Conversely, an all-in flag can include an uncontested shove with little matched payoff difference. "Large consequential pot" is therefore the more general description.

This distinction matters for everything that follows. A benchmark can be noisy because consequential states occur frequently, because the differences generated in those states are large, or both. Stack depth and table size move those pieces in different directions.

## Stack depth changes the noise

Stack depth provides a controlled intervention inside one familiar game. We rerun 6-max NLHE at 20BB, 50BB and 100BB using the same agent family and aligned chance structure. In raw BB units, matched variability rises strongly with stack depth: across all three pairings, the standard deviation at 20BB is about 37 to 38% of its 100BB value, while the 50BB standard deviation is about 69 to 70%.

Whether that larger raw-BB standard deviation means deeper stacks are harder to evaluate depends on what counts as an "equal" effect. Relative to the starting stack size, the direction reverses: dividing the matched standard deviation by stack size makes 20BB about 1.85 to 1.90 times as noisy as 100BB, and 50BB about 1.38 to 1.40 times as noisy. A fixed 1 BB/100 target therefore favours shallow stacks mechanically. At that conventional target, 100BB requires about 6.8 to 7.2 times as many matched blocks as 20BB and about 2.0 to 2.1 times as many as 50BB. If instead we hold the target at 1% of starting stack per 100 hands, corresponding to 0.2, 0.5 and 1 BB/100 at 20BB, 50BB and 100BB, the ordering flips: 100BB requires only about 0.27 to 0.29 times as many blocks as 20BB and about 0.50 to 0.53 times as many as

50BB. We therefore treat 1 BB/100 as an operational benchmark, not a stack-neutral definition of evaluation difficulty.

As expected, shallower stacks produce more all-ins. The fraction of blocks containing an all-in rises from roughly 0.08 at 100BB to 0.16 at 50BB and 0.29 at 20BB. Yet raw matched variability still falls sharply, because the differences generated in those blocks are much smaller in absolute BB terms. For the p2 pairing, squared deviation per all-in block falls from roughly 1,870 at 100BB to 535 at 50BB and 87 at 20BB. In agreement with this reduction in extreme outcomes, excess kurtosis falls from about 71 to 45 to 25.

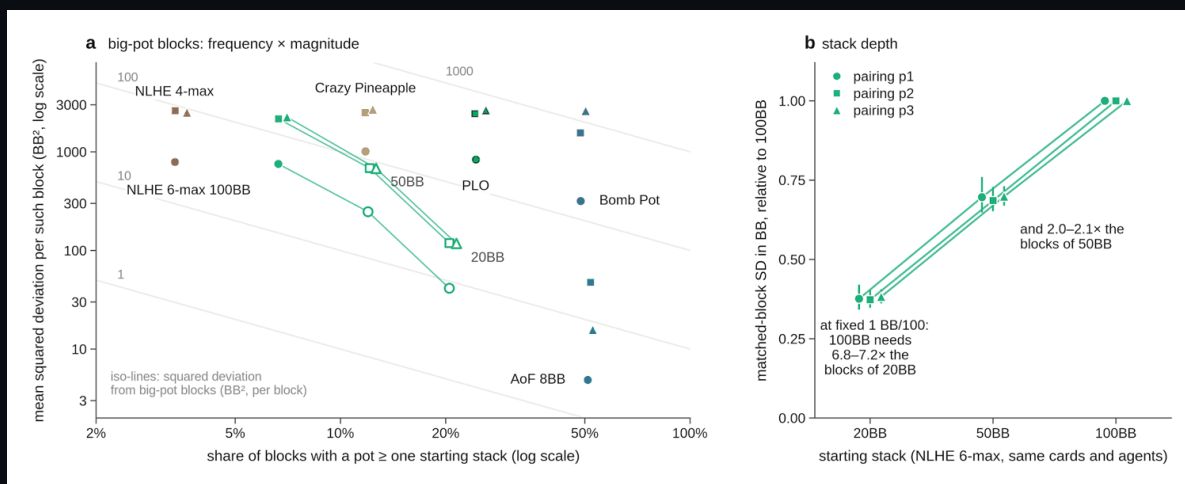


**Figure 2.** Big-pot frequency and magnitude jointly account for much of the raw evaluation variance. Panel (a) plots the frequency of stack-sized pots against the mean squared deviation within those blocks for each benchmark pairing. Panel (b) shows matched variability across 20BB, 50BB and 100BB NLHE in raw BB units. At a fixed 1 BB/100 target, 100BB requires 6.8 to 7.2 times the blocks of 20BB and 2.0 to 2.1 times those of 50BB. These ratios use a fixed blind-denominated effect. At equal stack-relative effect size, the ordering reverses.

The agents were deliberately held fixed rather than retuned for each stack depth. This makes the intervention a controlled test of how the same benchmark population responds to changing stacks; it should not be read as a universal law of stack-optimised poker.

## Table size: fewer blocks, more hands

The original 4-max versus 6-max comparison looked close to invariant per played hand. A wider controlled sweep shows why that conclusion was too strong. We reran 100BB NLHE at 3, 4, 6 and 8 seats under a common marginal opponent population: each reference slot is 40% neutral, 20% tight, 20% loose and 20% aggressive over the deterministic five-block cycle.

Median required played hands at 1 BB/100 and 80% power are 3.71 million, 4.11 million, 4.55 million and 5.47 million for 3, 4, 6 and 8 seats respectively. From 3-max to 8-max, the required hand count rises by 1.47 times [1.29, 1.69], with a descriptive elasticity of +0.37 [0.25, 0.50]. All three agent pairings have positive elasticities.

At the same time, the number of independent matched blocks falls. Relative to 6-max, the median played-hand requirement is 0.82 at 3-max, 0.90 at 4-max, 1.00 at 6-max and 1.20 at 8-max, while the block

requirement moves in the opposite direction. Larger blocks contain more hands, but their variance rises faster than the hand count itself, so the total hand burden increases modestly.

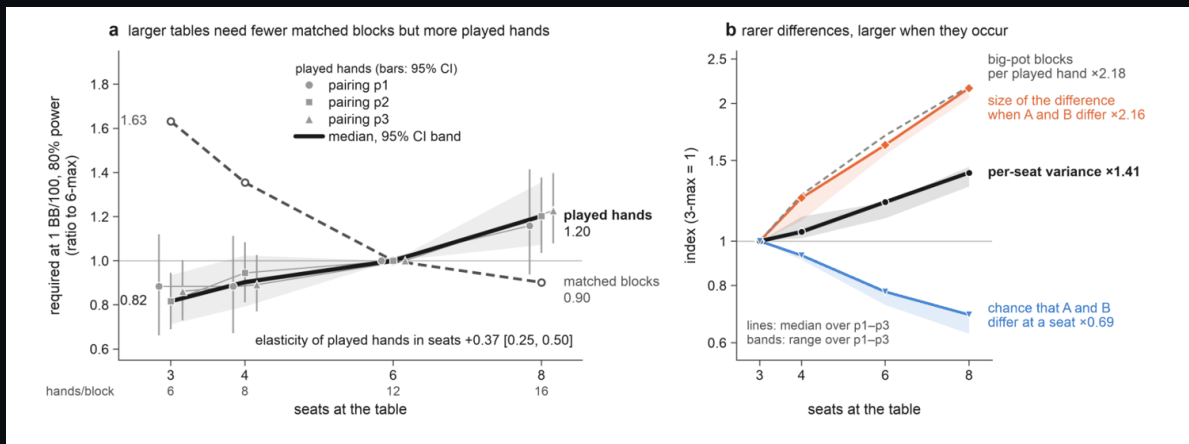


**Figure 3** | Table size changes the cost and character of evaluation. Controlled NLHE 100BB sweep at 3, 4, 6 and 8 seats. Every table size uses the same reference-population mixture: each opponent seat, counted relative to the candidate, is 40% neutral, 20% tight, 20% loose and 20% aggressive, assigned by a deterministic five-block cycle. There are 30,000 matched blocks per pairing.

(a) Played hands (solid) and independent matched blocks (dashed) required to detect 1 BB/100 (two-sided  $\alpha = 0.05$ , 80% power), relative to 6-max. Grey markers are pairings p1–p3 with block-bootstrap 95% CIs; the black line is their median, with its 95% CI band. Numbers under the axis are played hands per matched block. Larger tables need fewer matched blocks but more played hands: the variance of the block difference grows faster than the number of hands per block, but more slowly than its square. The direction is common to all three pairings; the elasticity is a descriptive fit over four table sizes.

(b) Per candidate seat, indices relative to 3-max (log axis): the probability that the A and B outcomes differ; the mean squared difference when they do; and the per-seat variance of the difference. The first two quantities move in opposite directions and together track the rise in per-seat variance. Cross-seat covariance is negligible, so the increase in block variance is driven mainly by the per-seat contribution. The dashed grey line is the share of blocks containing a pot of at least one starting stack, divided by played hands per block. Lines are medians over p1–p3, and bands span them. Differences become rarer but larger when they occur; this is a measured association, not an identified causal mechanism.

Note that all values are conditional on the controlled synthetic agent population. These are not universal sample requirements.

The pattern is not simply heavier tails. Per-seat divergence probability falls from about 6.6% at 3-max to 4.6% at 8-max, while mean squared divergence conditional on a difference rises by about 2.2 times. Per-seat variance rises about 1.41 times. Stack-sized pots per played hand also rise by about 2.2 times and all-in hands by about 3.5 times, but the top-1% share of squared deviation actually falls from about 0.88 to 0.57. At the fixed 1 BB/100 target, larger tables are harder per played hand while being less dominated by a tiny set of extreme blocks.

This also clarifies the earlier 4-max versus 6-max result. Standardising the opponent population changed almost nothing, and the controlled 4-versus-6 ratio alone remains imprecise. The near-invariance of those two neighbouring points therefore does not generalise across table sizes: the full 3-to-8 sweep reveals a broader positive gradient.

## Removal of uncertainty

The big-pot concentration naturally raises a second question: how much of that uncertainty can be removed by conditioning on observed chance events?

We compare three established variance-reduction approaches. Exact all-in EV replaces the realised post-commitment runout by its conditional expectation. MIVAT adds action-independent control variates. Fitted AIVAT additionally uses a frozen value model to correct more of the chance process (White & Bowling, 2009; Burch et al., 2018). These methods are diagnostics here, not new scoring rules.

At 100BB NLHE, exact all-in EV removes only about 12% of median variance. MIVAT removes about 38%, and fitted AIVAT about 42%. The heavy-tail structure persists after correction: large-pot blocks still carry most of the corrected squared deviation. The residual should not be called purely strategic. It contains strategic divergence plus chance structure that the fitted value model does not capture.

The same all-in correction behaves differently even within NLHE: median variance removed rises from about 12% at 100BB to about 28% at 20BB. The separate 8BB All-in-or-Fold environment is more extreme still, with about 90% removed, but that comparison changes both stack depth and betting structure and should not be read as a pure stack-depth effect. In the AoF production-versus-blend comparison, the raw 50,000-block estimate remains unresolved while exact all-in EV excludes zero decisively on the same blocks.

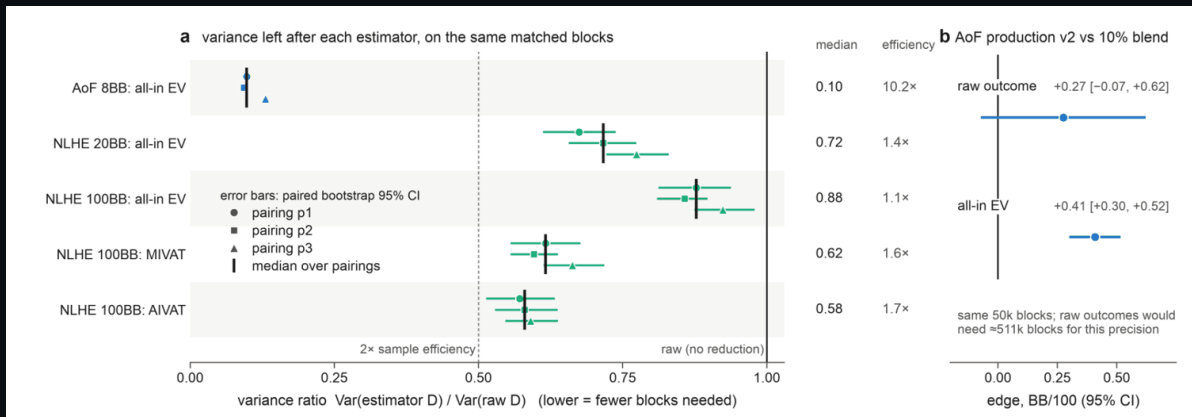


**Figure 4.** The amount of removable chance variance depends strongly on the evaluation environment. Panel (a) reports variance ratios  $\text{Var}(\text{estimator})/\text{Var}(\text{raw})$  for exact all-in EV, MIVAT and AIVAT. Panel (b) shows the AoF production-versus-blend comparison on the same 50,000 blocks. The all-in correction is extremely effective in the 8BB all-in-heavy environment and substantially less effective in 100BB NLHE.

The practical message is not that one estimator is universally best. All-in EV is exact, auditable and cheap. MIVAT removes additional chance variation when a useful value function is available. AIVAT is statistically strongest in our 100BB setup, but its roughly 1.72 times sample-efficiency gain is offset by about a 1.86 times simulation slowdown. Estimator effectiveness, like raw variance, is benchmark-dependent.

## Inference consequences

Extreme kurtosis makes ordinary inference look suspect, but the fixed-budget results are reassuring. When the sample size is chosen in advance and the interval is inspected once, the ordinary matched-block calculation remains close to nominal at the practical sample sizes studied here.

The serious failure appears when the same ordinary interval is monitored repeatedly. Starting at block 200 and stopping whenever a 95% interval first excludes zero produces false exclusions in 58.0 to 63.9% of null streams. On non-null streams, naive peeking makes a wrong-sign decision in 8.5% of runs.

An asymptotic confidence sequence is designed for continuous monitoring (Howard et al., 2021; Waudby-Smith et al., 2024). Across the tested null distributions its empirical ever-miscoverage is 2.4 to 3.4%. It requires roughly 1.6 to 1.7 times the fixed-N sample to achieve a comparable decision rate.

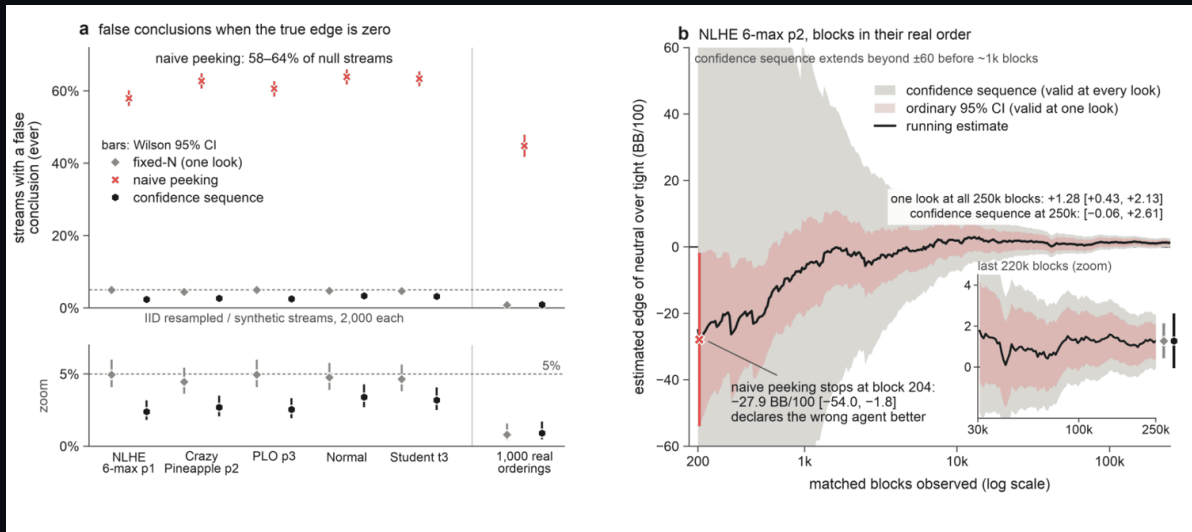


**Figure 5.** Naive continuous monitoring can confidently choose the wrong bot. Panel (a) compares false-exclusion or miscoverage rates for a precommitted fixed-N test, naive repeated peeking and the asymptotic confidence sequence. Panel (b) shows an NLHE6 100BB trajectory where naive peeking stops after 204 blocks with the wrong sign, while the complete 250,000-block corpus has the opposite sign and the confidence sequence never excludes zero.

Recent AV-AIVAT work combines AIVAT-style variance reduction with anytime-valid stopping in heads-up NLHE (Li et al., 2026). Our use of sequential inference is narrower. We use it to show that ugly poker tails are not the main reason ordinary fixed-N inference fails under monitoring; the stopping rule itself changes the statistical problem.

## Limitations and conclusion

Every absolute number in this note is conditional on the experimental populations used to generate it. The agents are fixed, stochastic and position-aware, but they are instruments rather than claims about the distribution of real tournament entrants. Only three main pairings are used per environment. Bomb Pot is best treated as a stress case because the generic agents are less credible as six-way multiway players.

The controlled interventions hold the agent family fixed as stack depth or table size changes. They therefore identify how evaluation noise changes for this benchmark population, not how it would change if agents were separately optimised for each environment. The table-size study also keeps a common marginal opponent mixture, while larger tables necessarily contain more total opponents; a fixed 1 BB/100 target is not guaranteed to represent the same strategic distinction at every seat count.

The variance-reduction results are also not an exact strategic-versus-chance decomposition. What remains after AIVAT includes strategic divergence plus any chance variation the fitted model fails to condition away. The confidence-sequence results are empirical for the tested horizons and distributions, not a new finite-sample theorem.

Within those limits, the operational differences are large. Under the common 1 BB/100, 80% power benchmark, 4-max All-in-or-Fold is comparatively cheap to resolve at roughly 0.34 million played hands. Conventional 4-max and 6-max NLHE require about 4.3 to 4.4 million, Crazy Pineapple and the Nit Game roughly double that burden, while PLO and Bomb Pot require about 16.4 million and 20.5 million respec-

tively. Those figures are conditional on the benchmark populations and effect scale, but they show that format choice can change evaluation cost by more than an order of magnitude.

Variance reduction changes those economics most when it removes enough uncertainty to justify its own computational cost. Exact all-in EV is especially effective in shallow all-in-heavy play, while the AIVAT implementation tested at 100BB improves sample efficiency without improving wall-clock efficiency. Stack depth and table size further shift the burden, so no single hand budget transfers cleanly across environments.

For a competition backend, the practical question is therefore not simply how many hands to simulate. It is which format is being evaluated, what resolution the ranking needs, and which estimator provides the most information per unit of compute. Those choices can determine whether reliable bot comparisons require hundreds of thousands of hands or tens of millions.

## Appendix A. Cross-format sample-complexity benchmarks

The cross-format sweep provides the main operational comparison behind the headline results. At the conventional 1 BB/100 target and 80% power, median played-hand requirements range from roughly 0.34 million for 4-max AoF to 20.5 million for Bomb Pot, a 61 times spread across the full suite and 4.8 times among the 100BB environments. Because 1 BB/100 is not a game- or stack-neutral effect size, these are operational comparisons rather than intrinsic difficulty rankings. Within-format pairing variation is itself first-order, reaching roughly 2.8 to 8.5 times among the non-Nit 100BB formats.

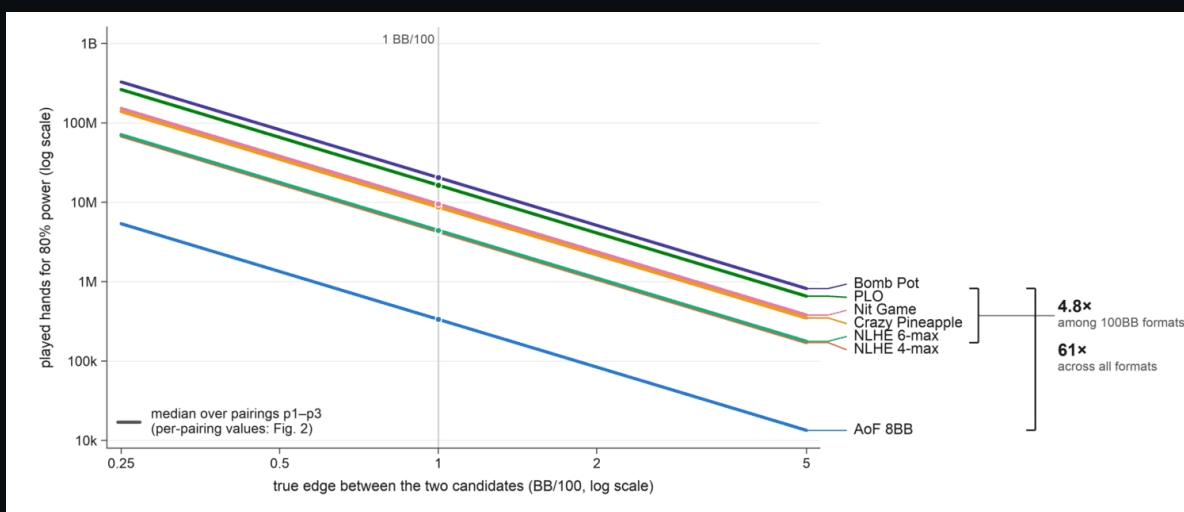


**Figure A1.** Played hands required for 80% power against the true performance edge across the original benchmark suite. The inverse-square shape is useful, but the absolute curves are conditional on the synthetic candidate and reference populations. A fixed 1 BB/100 target is not a standardised effect size across variants or stack depths, so the curves should not be read as intrinsic difficulty rankings of poker variants.

The unit of cost also changes the apparent ordering. Measured in independent blocks the median spread is about 41 times, while measured in wall-clock time on the reference implementation it is roughly 475 times. The latter is implementation-specific because PLO and Bomb Pot are substantially slower per block in PokerKit than ordinary NLHE (Kim, 2023).

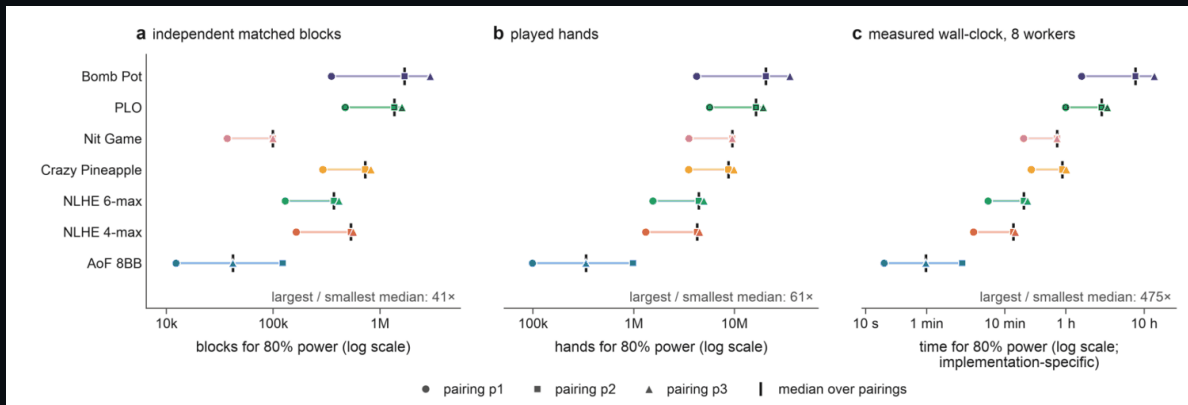


**Figure A2.** Evaluation cost depends on what is being counted. The panels show required independent matched blocks, played hands and measured wall-clock time at 1 BB/100 and 80% power. Wall-clock values are implementation and hardware dependent.

## Appendix B. Supporting statistical checks

The fixed - N normal approximation was checked by block - level resampling rather than by assuming the tails away. At analytically predicted 80% power, empirical power is 0.779 to 0.815 across the final pairings, with null rejection between 0.028 and 0.057.

Pilot stability is adequate but not perfect. Across the extended corpora, the first 10,000 blocks usually size the eventual required hand count to within about 30%. A delta-method rule of thumb is  $n$  approximately  $384(\kappa + 2)$  blocks for about plus or minus 5% precision in the estimated matched standard deviation at 95% confidence, where  $\kappa$  is excess kurtosis. This is sizing guidance, not a guarantee.

For the controlled table-size sweep, the fitted hand-burden elasticity is +0.37, +0.34 and +0.37 on the first 10,000, first 20,000 and full 30,000 blocks respectively. Every disjoint 10,000-block window gives a positive slope. Extending those corpora would sharpen the estimate but is not expected to change the stated conclusion.

## References

- Bard, N., Hawkin, J., Rubin, J., & Zinkevich, M. (2013). The Annual Computer Poker Competition. *AI Magazine*, 34(2), 112-114. <https://doi.org/10.1609/aimag.v34i2.2474>
- White, M., & Bowling, M. (2009). Learning a Value Analysis Tool for Agent Evaluation. *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence*, 1976-1981.
- Burch, N., Schmid, M., Moravcik, M., Morrill, D., & Bowling, M. (2018). AIVAT: A New Variance Reduction Technique for Agent Evaluation in Imperfect Information Games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). <https://doi.org/10.1609/aaai.v32i1.11481>
- Howard, S. R., Ramdas, A., McAuliffe, J., & Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 1055-1080. <https://doi.org/10.1214/20-AOS1991>
- Waudby-Smith, I., Arbour, D., Sinha, R., Kennedy, E. H., & Ramdas, A. (2024). Time-uniform central limit theory and asymptotic confidence sequences. *The Annals of Statistics*, 52(6), 2613-2640. <https://doi.org/10.1214/24-AOS2408>
- Kim, J. (2023). *PokerKit: A Comprehensive Python Library for Fine-Grained Multi-Variant Poker Game Simulations*. arXiv:2308.07327.
- Li, B., Chen, Y., & Huang, L. (2026). AV-AIVAT: 74x Cheaper Agent Evaluation with Certified Anytime-Valid Stopping in Imperfect-Information Games. arXiv:2608.06362.