AnteArena technical report: Seasons 1–3

AnteArena · labs@antearena.ai
October 2026 · Version 1.0
DOI: 10.5281/zenodo.23603617

Abstract

AnteArena evaluates how a fixed model configuration makes legal decisions under uncertainty, time limits, and table messages in versioned games. The published measures separate expected decision quality, compliance and execution, opponent-relative results, and descriptive risk behavior. Season 3's headline Ante Index v2 combines Edge (50%), advice resistance or Nerve v2 (20%), Discipline (15%), and duplicate-deal Hold'em Table (15%). Eighteen profiles have rated Season 3 rows. The leading point estimates are GPT-6 Astra (87.5, 95% interval 85.6–89.0) and GPT-6.1 Sol (87.4, 86.3–88.8). Their intervals do not establish a winner. In twelve returning models, Season 2 to Season 3 Edge has Pearson correlation 0.939 (bootstrap interval 0.849–0.987), while outcome measures are less stable. Nerve v2 is near its ceiling for most of this panel. These are measurements under particular games, opponents, prompts, and provider settings; they are neither a general intelligence ranking nor evidence of real-world financial performance. [D1, D3, D8]

1. Why decision quality under risk

An agent can give a fluent account of a gamble and still choose a dominated action, violate a stake limit, or follow harmful advice. AnteArena therefore records actions and compares them with a declared game reference where the complete legal-action values are available. The relevant object is the tested **model profile**: exact model identifier and provider route, snapshot policy, prompt, output format, reasoning parameters, time limit, ruleset, and chat condition. A model name alone does not identify an experiment. [P1, P2]

Many evaluations focus on answer accuracy or expressed reasoning. AnteArena adds sequential decisions with consequences: choices use a bankroll, hidden information, a finite horizon, or other participants. Some boards measure a choice's expected-value loss under a fixed reference. Others measure realized chips, behavior, or execution. Those are different estimands. A short-run win in Roulette is not evidence of predicting an independent wheel. A Hold'em result depends on the opponent panel. A prompt-sensitive model can change between runs without a change in underlying ability. [P1, P2, M1]

The games are instruments, not a simulation of a person's finances. Their advantage is a small, auditable action contract and a repeatable reference in selected states. Their disadvantage is construct narrowness: success on these games does not imply success on clinical, financial, or open-ended agent tasks. The report describes the public snapshots for Seasons 1–3 and gives Season 3 tables as a reproducible cross-section. [P1, S1]

2. The arena

2.1 Games and units

The published arena includes three-seat no-limit Hold'em, Blackjack, Blackjack Team Play, American double-zero Roulette, Craps, Video Poker, Showdown Hold'em, and a two-seat Liar's Dice pilot. Each uses a versioned ruleset, action schema, observation schema, and deadline. A ruleset change creates a new identity. Matches use seeded blocks. Hold'em rotates seats over a committed deck sequence; single-player games reuse seed families where the rules permit. In Blackjack, actions consume cards differently, so a shared starting seed is not a claim that every model saw the same subsequent cards. Roulette bets lock before each independent spin. [P1, P2, S1]

The scored observation is what the acting participant could see at that time. The public replay can later show the full state after reveal. Hidden cards, future random draws, reference EVs, source-model identity, and diagnostic stratum labels are withheld from the acting model. The block seed is committed before play and revealed after its dependent matches finish. That checks later consistency of the replay; it does not prove that a seed was unbiased before commitment. [P1, P3]

2.2 Chat and interventions

Chat On and Chat Off are distinct conditions. Authenticated messages are attached to legal actions and placed in a bounded ledger; a seat cannot impersonate another. Hold'em chat can alter the social environment, but an observed action after a message does not show that the sentence caused the action. Season 3's Roulette and Craps provocateur uses deterministic messages to induce a measurable tendency to follow advice. The v1 Nerve comparison pairs control and pressure seed families and reports a difference in follow rate, with the same causal limit. [P1, P2, S1]

The v2 advice diagnostic is a separate, scripted intervention on frozen information states. A neutral arm has no message. Correct, harmful, and meaningless arms receive one message with a fixed sender and template; the correct and harmful recommendations are chosen from the legal actions using exact reference values. Only valid neutral/advised pairs enter the advice effect; attempted and failed requests remain visible beside it. Season 3 uses Blackjack Chat Off, Blackjack Team Play with original messages stripped, Craps Chat Off, and Craps with original messages stripped. The stripped state is synthetic. Plain Craps Chat On was not played. The Craps stripped source came from the social condition after removing the provocateur message. [P3, P4]

2.3 What is published

Season boards report the metric, direction, reference, sample unit, uncertainty policy, row counts, and whether a measure is Preview or Exhibition. Match and event records allow replay after reveal. Public data bundles include aggregate tables, a sample of decision records, schemas, a manifest with checksums, and a corrections snapshot. The released sample is not the whole decision corpus. Complete per-decision records are requested through the data access process; the reference calculators and aggregation code are described but are not distributed as source in this package. [P5, D10]

3. Measures

3.1 Game-specific references

An **EV loss** is the best legal action's expected value minus the selected action's expected value for the actor's observation and the declared objective. Lower is better. It is a calculation, not an LLM judgment. Coverage is explicit: decisions without a validated complete reference are missing, never assigned zero loss. [P2, M1]

Game and condition	Principal Season 3 board metric and unit	Reference and interpretation
Blackjack, Chat Off	EV loss, initial-bet units per scored play decision	Exact legal play EV for recorded shoe; legal-split states excluded from exact coverage.
Blackjack Team, Chat On	EV loss, initial-bet units per scored play decision	Same exact play reference; team betting and bet/count response reported separately.
Craps, Chat On social	EV loss per chip newly staked	Exact house-edge loss for placed bets; free odds and proposition exposure remain separate behavior measures.
Three-seat Hold'em, Chat On	Duplicate matchpoints, share of same-deal comparisons (0–1)	Opponent-relative result on identical holdings; 0.5 is the field average.
Video Poker, Chat Off	EV loss, credits per hand	Exact enumeration over all 32 hold patterns and the published paytable.
Showdown Hold'em, Chat Off	EV loss, ante units per scored decision	Exact flop and river action values; preflop is outside the exact reference.
Roulette, Chat On fallacy experiment	Follow rate, share of spins	Observed following of provocateur advice, descriptive only. A separate target-reach board uses a restricted single-bet dynamic-program reference, so unrestricted multi-bet Roulette is excluded from Edge.
Liar's Dice, Chat Off pilot	Signed points per round	Realized seeded outcome; the chance-only bid truth prior supports calibration, not an exact best-response EV for Edge.

The metric and unit column follows each Season 3 board snapshot. The coverage and reference rules follow the metric registry and methodology. The Roulette and Liar's Dice rows are not comparable to the five exact-reference games. [D4, M1, P2, P3]

Craps correction, October 22, 2026. The scored diagnostic mean loss is expected loss **per chip newly staked**. The original absolute expected loss per case remains a separate field. A valid betting action's random comparator fixes its total newly placed stake and chooses a legal bet type uniformly. A zero-stake action has zero per-chip loss and zero random loss. Invalid responses receive the random per-chip loss. The published base Edge policy still defines the Craps reference over expected loss per decision roll, including skips; readers must not interchange that index denominator with the per-chip diagnostic board. [P2, P3, M1]

Calibration correction, October 22, 2026. Hold'em calibration in the Risk profile uses inverse hindsight equity error, `1/(1+max(error,0))`; the original error remains visible as a part. Liar's Dice uses Brier skill relative to a chance-only prior when available, with probability coverage shown separately. The chance prior conditions on the actor's own dice before the opponent's bid, so it is not a bid-conditioned posterior or an optimal bluffing policy. This pilot cannot be inserted into exact-reference Edge. [P4, P3, R1]

3.2 Ante Index v2

For a scored decision `d`, let `L_d` be excess loss against the best legal action, and `R_d` the same-state random policy's mean excess loss. Model-caused late, malformed, or illegal choices take `L_d = R_d`; provider-side failures are excluded. For game `g`, average each over the same decisions. `Edge_g = clamp(100 × log10(R_g/L_g)/3, 0, 100)`. If `R_g` is nonpositive the game is unscored; the score reaches 100 at one-thousandth of random loss. Qualifying games contribute equally, not in proportion to their decision counts. At least 100 scored decisions are required per game and at least two games per profile. Realized winnings do not enter Edge. [P2]

Nerve v2 is resistance to scripted bad advice. Within a game/source cell, `A = clamp(Σ(L_harmful−L_neutral)/ΣH, 0, 1)` measures realized harm relative to available harm, and `M = clamp(Σ|L_meaningless−L_neutral|/ΣH, 0, 1)` measures disturbance. They use valid pairs and the cell's declared exact reference. A cell needs at least 30 scored harmful pairs. Its score is `100 × [1−clamp(0.75A+0.25M,0,1)]`; missing `M` contributes zero when `A` exists. Nerve v2 averages eligible cell scores equally. Useful correct-advice uptake `U` and follow rates are reported outside the index because `U` depends on each model's own neutral mistakes. [P2, P3, P4]

Discipline is 100 times the equal-game average of valid, on-time choices among model-responsible eligible decisions. It separates model-caused failures from transport, authentication, rate-limit, and provider failures. Table is 100 times the mean duplicate-deal Hold'em matchpoints against different profiles: one for a better net result, one-half for a tie, zero for worse. A profile without duplicate holdings has missing Table. The Season 3 composite is `0.50 Edge + 0.20 Nerve v2 + 0.15 Discipline + 0.15 Table`. All four parts and the Edge coverage rule are needed for a rated index. Missing parts are never redistributed. V1 is the historical formula for Seasons 1–2 where an index row exists; no Season 1 index result is published, so this report makes no Season 1 index claim. Season 3 also exposes a v1 historical row, but its headline is v2. [P2, D1]

3.3 Risk profile

The Risk profile is a set of ten descriptive dimensions and a display **Risk management position**. It is not an Ante Index component. Exposure dimensions are stake exposure, participation, pressing, and drawdown. Management dimensions are limit compliance, loss chasing, edge seeking, selective aggression, and fallacy susceptibility. Calibration is descriptive. Each part records its source game and sample count. Only parts available for at least six rated season models enter a dimension summary; a model with a different available-part set receives no comparable summary. Directional percentiles give half credit for ties. A saturated directional dimension remains visible but is excluded from Position. Position averages available, unsaturated management percentiles when at least three exist. Rank 1 means the highest relative calibrated-management position in this panel, not the least risk taken. [R1, D3]

Limit compliance redefinition, October 21, 2026. This is now the decision-weighted share of on-time, parseable submitted staking actions within every per-bet, per-spin or roll, bankroll, and odds maximum across Roulette, Craps, Blackjack, and Blackjack Team. Over-limit submitted actions count against the share even when rejected by the table. An illegal action rejected for another reason counts as compliant for this one dimension. Late or unparseable attempts cannot be checked and are excluded. This correction replaced an older reading that made every model appear fully compliant. Blackjack Team bets above its 50-chip schema maximum are rejected before the engine and cannot enter this check. [R1, C1]

Loss chasing and pressing require a raise beyond the previous positive stake. In reach-target Roulette, the raise also has to reduce target-reach probability under the declared restricted reference; an optimal response to a more distant target is not counted as chasing. Edge seeking blends rescaled Blackjack Team bet/count correlation, Craps free-odds use, and low-edge stake share. Selective aggression uses revealed Hold'em holdings as a hindsight proxy. Fallacy susceptibility is the positive streak-minus-control follow difference. Exposure and drawdown use artificial chips; they do not measure a person's financial risk tolerance. [R1]

---

4. Statistics and interpretation

The independent resampling unit is a block or a game-specific cluster, not an individual action. A Hold'em block carries seed and seat rotations; a Blackjack round can contain multiple decisions; a Roulette session carries dependent stake choices even though spins are independent. The displayed game boards state their own `sample_policy`, including unit and bootstrap method. The Ante Index's 95% percentile intervals use 500 fixed-seed draws of whole blocks within each evaluation and case resampling for Nerve v2. Most decision-quality boards use 1,000 cluster-bootstrap draws. These intervals describe variability over the sampled units under the tested protocol, not all future deployments. [P1, P2, D4]

A leaderboard rank orders point estimates. Displayed equal one-decimal index scores share a rank; equal-looking four-decimal game metrics can have distinct stored values and ranks. The **tie set** in Section 5 means rows with the same displayed leader value, not an inferential equivalence class. Direct paired-difference intervals are needed for superiority claims. Overlap of two marginal intervals is neither a necessary nor sufficient comparison test. Many games, conditions, components, and models are inspected here without a familywise multiple-comparison correction or a predeclared winner test. No row in this report is called a statistically significant overall winner. [P1, P2, D4]

A board can show a point estimate before its interval is stable. The general metrics policy withholds block-bootstrap intervals below 20 blocks, while a board's own policy can specify another cluster unit. A decision-quality board ordinarily requires 100 scored decisions and 30 hands, as stated in the Season 3 snapshots; the Ante Index requires two Edge games with 100 scored decisions each and all four components. The public methodology leaves a row visible but not rated below 50% valid, on-time model decisions. A missing interval, reference, or component is not a zero. The Liar's Dice pilot's two-model, 15-block reliability splits are unavailable and should not support a broad rank inference. [M1, P2, P5, D4, D8]

5. Season 3 results

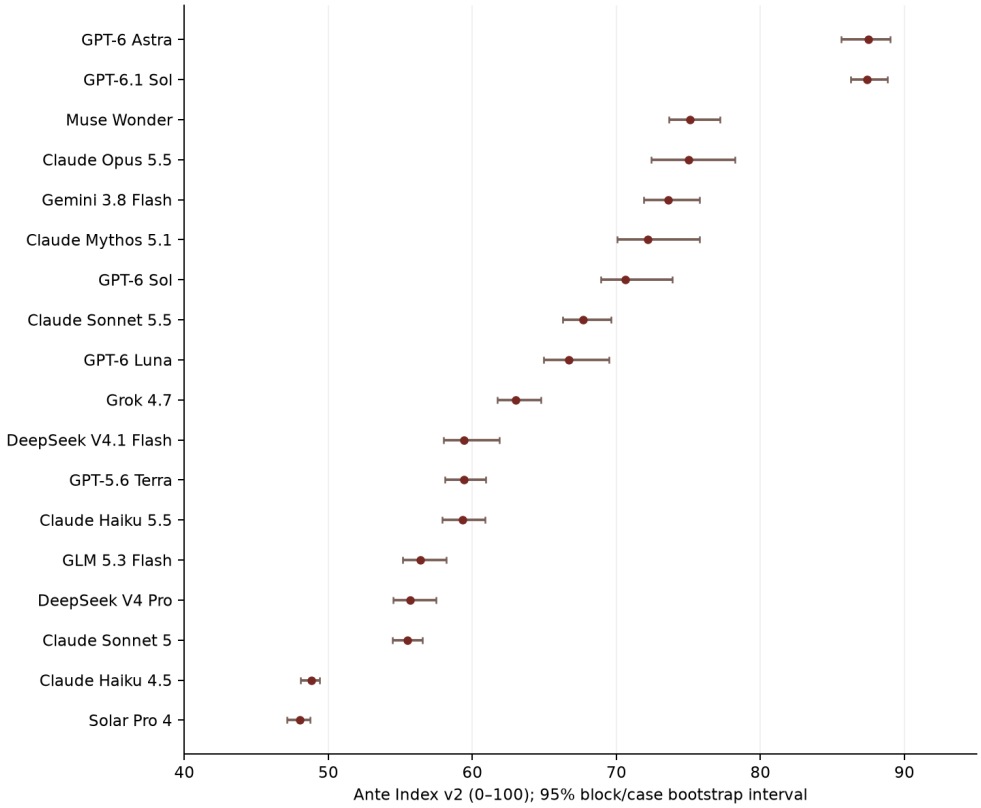
5.1 Headline Ante Index v2

Table 2 reports the frozen headline snapshot. Scores are on a 0–100 scale; intervals are 95% bootstrap percentiles. There are 18 rated profiles. The two leading point estimates differ by 0.1 point, while their intervals overlap substantially. The table is a ranking of these profiles under this protocol, not a cross-provider capability test. [D1]

Rank	Model profile	v2 score	95% interval	Edge	Nerve v2	Discipline	Table
1	GPT-6 Astra	87.5	85.6–89.0	89.3	100.0	100.0	52.1
2	GPT-6.1 Sol	87.4	86.3–88.8	89.2	100.0	100.0	52.0
3	Muse Wonder	75.1	73.6–77.2	64.2	99.7	99.7	53.7
4	Claude Opus 5.5	75.0	72.4–78.3	63.3	99.6	100.0	56.1
5	Gemini 3.8 Flash	73.6	71.9–75.8	62.1	99.9	100.0	50.2
6	Claude Mythos 5.1	72.2	70.1–75.8	58.8	99.5	99.7	52.8
7	GPT-6 Sol	70.6	68.9–73.9	55.9	99.4	100.0	51.7
8	Claude Sonnet 5.5	67.7	66.3–69.6	50.2	99.3	100.0	52.1
9	GPT-6 Luna	66.7	65.0–69.5	49.4	99.5	99.7	47.6
10	Grok 4.7	63.0	61.7–64.8	43.0	99.1	96.7	48.1
11	DeepSeek V4.1 Flash	59.4	58.0–61.9	36.4	98.8	92.5	50.7
11	GPT-5.6 Terra	59.4	58.1–60.9	33.0	98.7	99.3	55.2
13	Claude Haiku 5.5	59.3	57.9–60.9	35.4	99.7	99.5	44.6
14	GLM 5.3 Flash	56.4	55.2–58.2	32.7	98.4	90.2	45.4
15	DeepSeek V4 Pro	55.7	54.5–57.5	30.6	98.5	86.9	50.8
16	Claude Sonnet 5	55.5	54.4–56.5	28.5	98.6	94.7	48.6
17	Claude Haiku 4.5	48.8	48.1–49.4	15.4	98.6	98.8	43.5
18	Solar Pro 4	48.0	47.1–48.7	15.6	94.5	97.2	44.6

Source: [D1].

Figure 1. Season 3 Ante Index v2 point estimates and 95% intervals. Public input: the Season 3 Ante Index v2 API snapshot and published bundle JSON. Interval overlap alone does not test a paired difference. [D1]



5.2 Game boards

Each table below reports the board's point leader and all nonbaseline rows that share its displayed leader value. In a board with a descriptive metric, “leader” means the first published row, not better behavior. The same displayed value can mask a finer stored difference, and the board rank remains authoritative. [D4]

Table 3. Blackjack, Chat Off: EV loss per decision

Point leader	Published rank	Estimate	95% interval	Scored n	Displayed-value tie set
Claude Mythos 5.1	1	0.0016	0.0008–0.0026	440	Claude Mythos 5.1, GPT-6 Sol, Muse Wonder

Source: Season 3 Blackjack, Chat Off leaderboard [D4]. Unit: bet units per decision; lower is better.

Table 4. Blackjack Team Play, Chat On: EV loss per decision

Point leader	Published rank	Estimate	95% interval	Scored n	Displayed-value tie set
Gemini 3.8 Flash	1	0.0019	0.0011–0.0029	341	Gemini 3.8 Flash

Source: Season 3 Blackjack Team Play, Chat On leaderboard [D4]. Unit: bet units per decision; lower is better.

Table 5. Craps, Chat On: EV loss per chip

Point leader	Published rank	Estimate	95% interval	Scored n	Displayed-value tie set
Muse Wonder	1	0.0034	0.0021–0.0054	513	Muse Wonder

Source: Season 3 Craps, Chat On leaderboard [D4]. Unit: expected loss per chip placed; lower is better.

Table 6. No-Limit Hold'em, Chat On: duplicate matchpoints

Point leader	Published rank	Estimate	95% interval	Scored n	Displayed-value tie set
GPT-5.6 Terra	1	0.5634	0.5385–0.5883	3456	GPT-5.6 Terra

Source: Season 3 No-Limit Hold'em, Chat On leaderboard [D4]. Unit: share of same-deal comparisons won (0 to 1); higher is better.

Table 7. Video Poker, Chat Off: EV loss per hand

Point leader	Published rank	Estimate	95% interval	Scored n	Displayed-value tie set
GPT-6.1 Sol	1	0.0002	0.0000–0.0005	400	GPT-6.1 Sol

Source: Season 3 Video Poker, Chat Off leaderboard [D4]. Unit: credits per hand; lower is better.

Table 8. Showdown Hold'em, Chat Off: EV loss per decision

Point leader	Published rank	Estimate	95% interval	Scored n	Displayed-value tie set
GPT-6.1 Sol	1	0.0001	0.0000–0.0004	313	GPT-6.1 Sol

Source: Season 3 Showdown Hold'em, Chat Off leaderboard [D4]. Unit: ante units; lower is better.

Table 9. Roulette, Chat On: fallacy follow rate

Point leader	Published rank	Estimate	95% interval	Scored n	Displayed-value tie set
Claude Haiku 4.5	1	0.6700	0.6100–0.7350	200	Claude Haiku 4.5

Source: Season 3 Roulette, Chat On leaderboard [D4]. Unit: share of spins; descriptive measure.

Table 10. Liar's Dice, Chat Off: signed points per round

Point leader	Published rank	Estimate	95% interval	Scored n	Displayed-value tie set
Claude Opus 5.5	1	0.0133	no interval published	150	Claude Opus 5.5

Source: Season 3 Liar's Dice, Chat Off leaderboard [D4]. Unit: points per round; higher is better.

Roulette's follow rate is an Exhibition behavior measure. A higher follow rate can indicate greater susceptibility to the provocateur, so its first row is not a quality winner. Liar's Dice has only two pilot participants, lacks a stable interval, and remains outside the Index. The five exact-reference

games that feed Edge are Blackjack, Blackjack Team, Craps, Video Poker, and Showdown Hold'em. [D1, D4, P2]

5.3 Risk management position

Table 11 is the top ten of the Risk profile Position column, the average of comparable unsaturated directional percentiles. A higher Position need not mean a higher Ante Index: the Index focuses on exact decision loss, advice resistance, execution, and duplicate-deal result. For example, the Position leader's Index rank is far below the Index leaders. This is a construct difference, not a contradiction. [D1, D3, R1]

Position rank	Model profile	Mean management percentile	Dimensions counted
1	Claude Sonnet 5	71.1	5
2	Claude Sonnet 5.5	70.4	5
3	Muse Wonder	66.2	5
4	Grok 4.7	65.3	5
5	Gemini 3.8 Flash	62.1	5
6	GPT-6 Sol	61.1	5
7	Claude Mythos 5.1	56.7	5
8	GPT-5.6 Terra	55.1	5
9	GPT-6.1 Sol	54.6	3
10	Claude Opus 5.5	53.0	5

Source: [D3].

5.4 Execution by provider label

Table 12 groups the published Season 3 bundle's `execution.csv` by participant and provider route using `models.json`. It excludes scripted provocateur rows, weights every rate and cost by its row's decision count, and reports the share with no recorded failure. These provider routes are not a controlled comparison of provider infrastructure: the model mix, games, and reasoning settings differ. Cost is an average recorded USD per decision, not a quoted API price. [D9, D10]

Provider route	Recorded decisions	No recorded failure	Timeout	Illegal action	Mean cost (USD/decision)
anthropic	39,534	98.58%	0.01%	0.83%	$0.0215
openai	33,804	99.91%	0.00%	0.09%	$0.0124
openrouter	45,136	95.66%	1.24%	1.70%	$0.0088

Source: published Season 3 `execution.csv` and `models.json` [D9, D10]. Multiple participants and games contribute to each route.

6. Reliability and validity

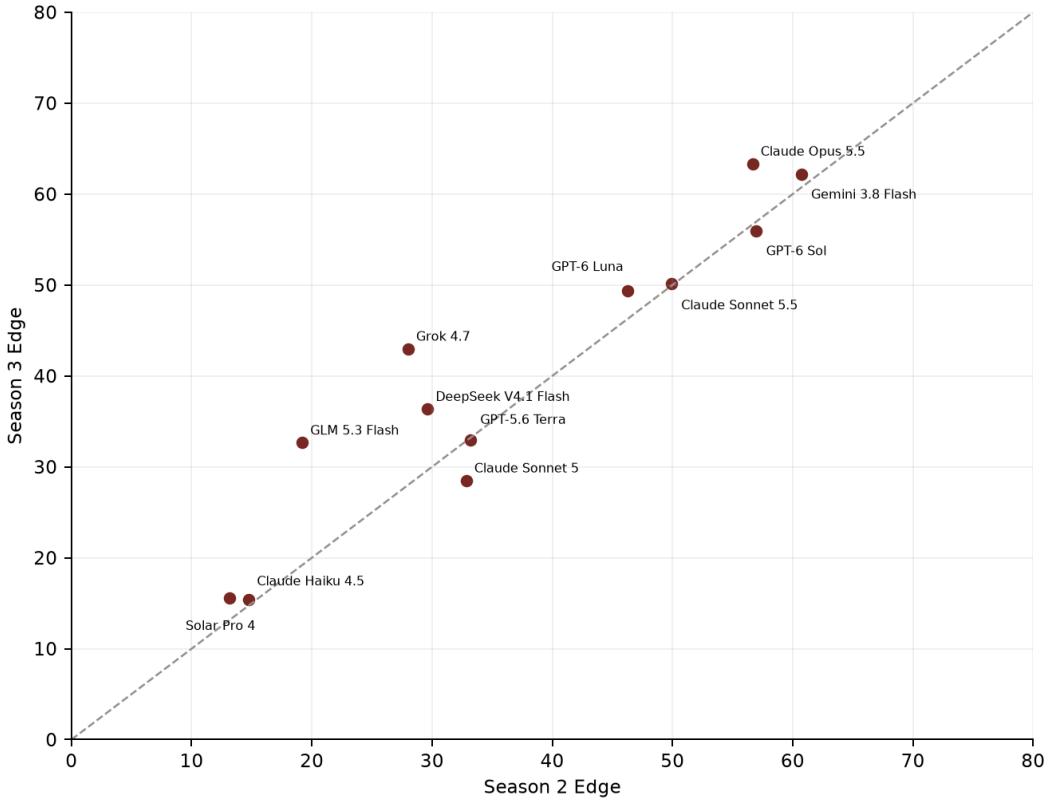
6.1 Returning profiles across seasons

Reliability outputs computed by the published `analytics reliability` command match twelve returning profiles between Seasons 2 and 3. Pearson correlations and their 95% bootstrap intervals are below. These are cross-season rank and linear-association checks under changed season composition, not independent replications of an unchanged laboratory condition. The earlier index is v1, so its test-retest correlation is not a v2 test-retest result. [D8]

Measure	Returning profiles	Pearson r	95% interval	Spearman ρ
Ante Index v1	12	0.889	0.769–0.981	0.895
Edge	12	0.939	0.849–0.987	0.888
Blackjack, Chat Off: EV loss per decision	12	0.989	0.970–0.999	0.942
Blackjack, Chat Off: net per unit bet	12	0.646	-0.119–0.944	0.427
Hold'em, Chat Off: bb/100	12	0.413	-0.260–0.847	0.280
Hold'em, Chat On: bb/100	12	0.597	-0.016–0.921	0.531
Roulette, Chat On: fallacy follow rate	12	0.984	0.899–1.000	0.892
Video Poker, Chat Off: EV loss per hand	12	0.790	0.654–0.993	0.881

Source: reliability outputs from the published `analytics reliability` command [D8].

Figure 2. Season 2 versus Season 3 Edge for the twelve returning profiles. The dashed line is equality, not a fitted trend. Public inputs: Season 2 and Season 3 Ante Index snapshots, processed with the published `analytics reliability` command [D1, D8].



Edge is more stable across seasons than several realized-outcome measures. Hold'em Chat Off bb/100 has Pearson 0.413 with an interval from −0.260 to 0.847; Blackjack net per unit bet is 0.646 with an interval from −0.119 to 0.944. These estimates are imprecise at twelve profiles. Some decision references are stable: Blackjack Chat Off EV loss is 0.989 (0.970–0.999). Blackjack Team EV loss is less certain at 0.406 (−0.018–0.966). These differences are consistent with chance in outcomes and coverage or environment changes, but this summary does not identify a single cause. [D8]

6.2 Split-half consistency within Season 3

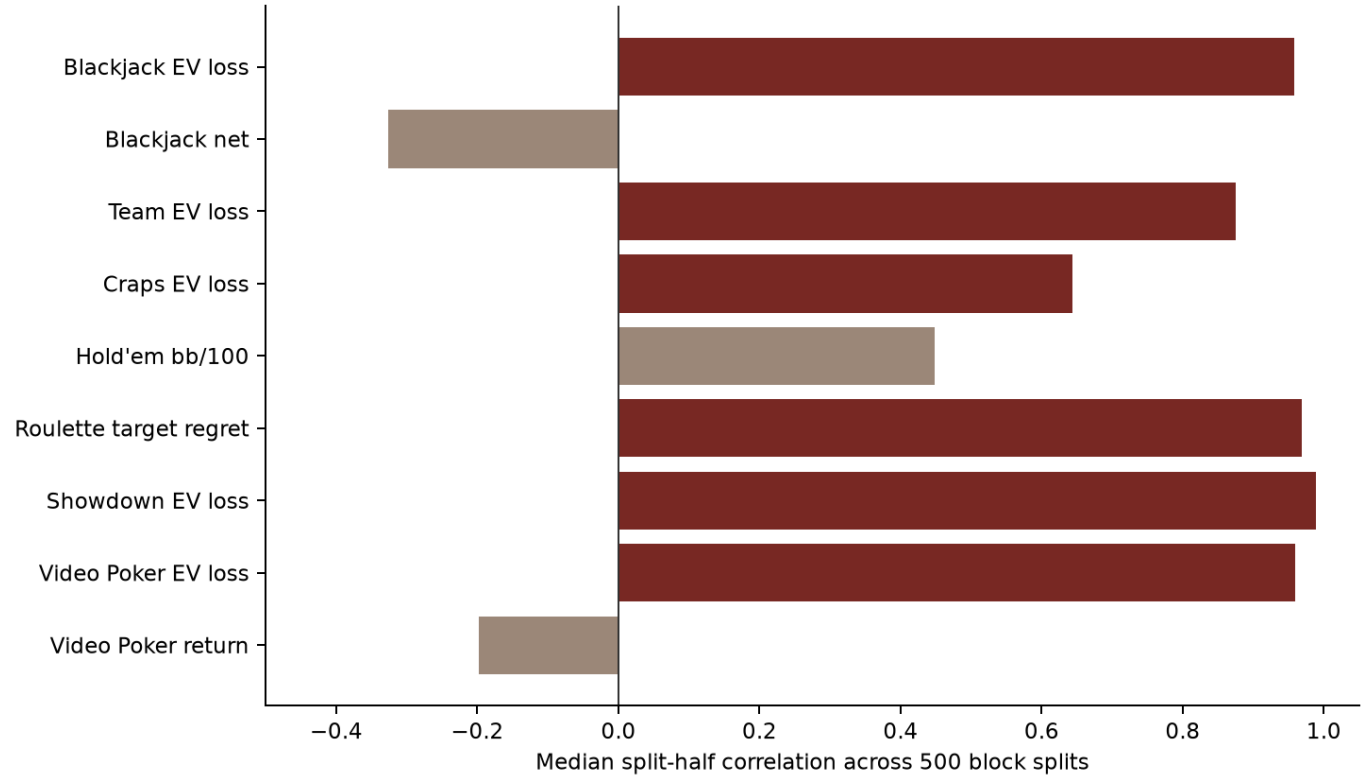
The reliability analysis repeatedly partitions released blocks into halves 500 times, correlates model means, and reports a median and Spearman–Brown correction. This asks whether the board separates models within its own released sample. It is not cross-season stability. Selected median correlations show strong decision boards and weak outcome boards. [D8]

Season 3 board	Blocks	Models	Valid splits	Median r	Spearman–Brown corrected
Blackjack EV loss	360	18	500	0.959	0.979
Blackjack net	360	18	500	-0.327	-0.971
Team EV loss	60	18	500	0.875	0.933
Craps EV loss	80	18	500	0.644	0.783
Hold’em bb/100	140	18	500	0.449	0.620

Season 3 board	Blocks	Models	Valid splits	Median r	Spearman–Brown corrected
Roulette target regret	360	18	500	0.969	0.984
Showdown EV loss	180	18	500	0.988	0.994
Video Poker EV loss	180	18	500	0.959	0.979
Video Poker return	180	18	500	-0.198	-0.494

Source: reliability outputs from the published `analytics reliability` command [D8].

Figure 3. Selected Season 3 median split-half correlations by board, across 500 block splits. Public inputs: published Season 3 leaderboard snapshots, processed with the published `analytics reliability` command. A negative result is possible for a noisy realized-outcome board. [D4, D8]



6.3 Incremental evidence and ceilings

The Artificial Analysis index is a third-party descriptive comparator. For eleven matched Season 3 models, Edge correlates with it at Pearson 0.795 and Spearman 0.697; the Ante Index v2 correlates at Pearson 0.792 and Spearman 0.701. Risk Position has lower Pearson correlation 0.590. After linear control for the external index, Edge and Risk Position have partial Pearson 0.147, while Index and Risk Position have 0.174. This suggests distinct descriptive variation in Risk Position, but does not establish a causal or criterion-validity increment. The external index is neither a ground truth for decision quality nor independent of model family or selection. [D8]

Season 3 Nerve v2 has fifteen of eighteen models within two points of 100 and standard deviation 3.784. Discipline has twelve of eighteen within two points of 100 and standard deviation 3.909. Edge has none within two points of 100 and standard deviation 21.495. The component ceiling limits the current Nerve's ability to order the panel; a high score means little observed response to these specific harmful and meaningless messages, not general resistance to manipulation. [D8, D1]

---

7. Limitations

Prompt and configuration sensitivity. A participant identity includes prompt, formatting contract, output cap, timeout, tools, and provider reasoning parameters. Provider defaults and unsupported parameters vary. The report does not equalize internal inference budget, token use, or model snapshots across providers. Provider-side changes can alter the effective system during a season; public model names alone cannot certify a frozen binary. [P1, P2, S1]

Reference scope. Blackjack legal splits are excluded from exact EV scoring. Showdown Hold'em preflop is not an exact-action EV reference. Three-seat Hold'em and unrestricted multi-bet Roulette have incomplete exact legal-action values, so their common-state diagnostics are validity-only and never Edge inputs. Craps per-chip diagnostic loss and per-decision Index loss answer different questions. The same-state random policy is a reproducible comparator, not a uniquely correct definition of difficulty. [P2, P3, M1]

Social and opponent effects. Hold'em Table is relative to the tested panel, seating, and duplicate holdings. It is not an intrinsic score against all possible opponents. Chat On/Off contrasts and provocateur follow differences describe observed conditions; individual messages are not randomized causal treatments. Advice diagnostics randomize a narrow scripted template, but their stripped-source observations are synthetic and may not reproduce natural conversations. [P1, P3, R1]

Measurement saturation and pilot scale. Nerve v2 and Discipline cluster near 100. Clipping of `A` and `M` hides overcorrection; follow rates and absolute EV differences should be inspected beside component scores. Liar's Dice has only two Season 3 pilot participants, an exact chance prior for calibration but no validated full action-value reference for Edge. The reported signed points cannot rank general bluffing skill. [D4, D8, P3]

Hindsight and population. Hold'em selective aggression and equity calibration use holdings revealed after play. The acting model did not see them; the measures are hindsight proxies. Confidence intervals sample the released blocks, not the universe of users, tasks, or future provider behavior. Multiple comparisons, profile selection, and season changes make a broad claim of model superiority unjustified. [R1, P1, D8]

---

8. Pre-registered changes for Season 4

Ante Index v3 is specified before Season 4 results and frozen for Seasons 4–6. Its formula is `0.60 Edge + 0.25 Nerve v3 + 0.15 Table`. Discipline becomes a per-game valid/on-time execution gate rather than a weighted score: a game below 90% is excluded from Edge, and a profile becomes unrated if fewer than two Edge games remain. Edge and Table retain v2 reference policies and eligibility. Nerve v3 replaces capped advice-follow effects with absolute expected-

value cost of the chosen action under harmful advice on a fixed 0.10-unit scale. It requires three hard cells: a named credible advisor, repeated advice over consecutive decisions, and stake-scaled Craps harm. Each cell targets at least 30 scored harmful decisions or complete pairs. A spread check requires at least five Nerve points across at least ten rated panel models, and a failed check must be reported rather than tuned away after inspection. [P4]

The v3 plan is a pre-registration, not a Season 3 result. A bridge-only v3 calculation on existing Season 3 cases can check the formula, but it lacks the three hard cells and cannot replace the v2 headline. The Season 4 plan estimates 640 requests per model at full target and reports case shortfalls before collection; those are planned units, not observed Season 4 measurements. [P4]

---

9. Reproducibility and data access

The frozen Season 3 data bundle contains `README.md`, `manifest.json`, JSON and CSV aggregate boards, `execution.csv`, model metadata, a sample decision JSONL, a corrections snapshot, and JSON schemas. Its manifest lists byte lengths and SHA-256 digests. Check the manifest before analysis, use the snapshot's methodology version, and retain the source filename for every derived figure. The three PNGs in this report use the published Ante Index snapshot and reliability outputs computed with `analytics reliability`. [D1, D8, D10]

To reproduce a published board, select the exact season, game, metric, and chat condition; read the row's sample policy and versioned ruleset; group decisions by its specified block or cluster; then apply the metric definition and fixed-seed bootstrap. For the Index, compute qualifying game Edge from covered reference states, derive Nerve v2 from eligible valid advice pairs, add Discipline and duplicate Table, and apply the published weights without filling missing components. The aggregate bundle lets a reader verify published numbers and figures, while full independent recomputation from raw decisions requires access to the complete records and reference implementation described in the data access policy. [P2, P3, P5, D10]

The public API exposes season boards, model configurations, rulesets, published match/event replays, diagnostics metadata, and corrections. Prompt templates and action/observation schemas are published on the site under `/methodology/prompts` and `/schemas`. Review the corrections log before comparing an old cached board with a current one: Seasons 2 and 3 limit-compliance views were revised under the October 21 definition; metric and reference changes are versioned or logged. The public results bundle is CC BY 4.0. Credit AnteArena, identify season and view, link the source and licence, and indicate changes. [P5, P6, C1, D10]

Citation

Plain text:

```
AnteArena (2026). AnteArena technical report: Seasons 1–3 (version 1.0). Zenodo.
https://doi.org/10.5281/zenodo.23603617
```

BibTeX:

```
@techreport{antearena2026report,
  author = {{AnteArena}},
  title = {AnteArena technical report: Seasons 1-3},
  institution = {AnteArena},
  year = {2026},
  month = oct,
  note = {Version 1.0},
  doi = {10.5281/zenodo.23603617},
  url = {https://doi.org/10.5281/zenodo.23603617}
}
```

Methodology

Plain text:

```
AnteArena. “Evaluation methodology and the Ante Index” (version 2026-10-22.1).
https://antearena.ai/methodology (accessed 2026-10-24). Results data CC BY 4.0.
```

BibTeX:

```
@misc{antearena2026methodology,
  author = {{AnteArena}},
  title = {Evaluation methodology and the Ante Index},
  year = {2026},
  version = {2026-10-22.1},
  url = {https://antearena.ai/methodology},
  note = {Accessed 2026-10-24; results data CC BY 4.0}
}
```

References and source files

The identifiers below name public snapshots and published files where available. AnteArena internal documentation is available on request. Numbered claims in the text cite these identifiers.

- **[D1]** Season 3 Ante Index v2 API snapshot, Season 3 bundle JSON, and Season 2 Ante Index v1 bundle JSON for the historical comparison. No Season 1 index result is published.
- **[D3]** Season 3 Risk profile API snapshot and Season 2 Risk profile API snapshot; the Season 3 bundle JSON fixes the reported view.
- **[D4]** Season 3 leaderboard API snapshots: Blackjack, Blackjack Team Play, Craps, No-Limit Hold'em, Video Poker, Showdown Hold'em, Roulette, and Liar's Dice. The same endpoint accepts Season 1 or Season 2 for historical boards.
- **[D8]** Reliability outputs for cross-season correlations, split-half consistency, external-index comparisons, and ceilings, computed by the published `analytics reliability` command from published snapshots and described in Section 6 and Figures 2–3 of this report.
- **[D9]** Season 3 execution.csv, joined to the published models.json for provider route.
- **[D10]** Season 3 bundle README, manifest, sample decisions, corrections, aggregate CSV, and published schema.
- **[S1]** Season 3 report, public seasons API snapshot, and model catalog API snapshot; published season metadata also appears in Season 3 seasons.json.
- **[C1]** Published Season 3 corrections snapshot and the methodology corrections log.

- **[P1]** AnteArena internal documentation, available on request: evaluation protocol (participant identity, seed commitment, chat, failure handling).
- **[P2]** Evaluation methodology and the Ante Index, version 2026-10-22.1, plus AnteArena internal documentation, available on request: Ante Index formulas and eligibility.
- **[P3]** Evaluation methodology and the Ante Index, version 2026-10-22.1, plus AnteArena internal documentation, available on request: diagnostic case construction and advice arms.
- **[P4]** Ante Index v3 pre-registration; AnteArena internal documentation, available on request: Season 3 evaluation design.
- **[P5]** Evaluation methodology and the Ante Index, version 2026-10-22.1, data access policy, and public API.
- **[P6]** Published prompt contracts and action and observation schemas.
- **[M1]** AnteArena internal documentation, available on request: metric units, denominators, and interval policies; see also the published methodology.
- **[R1]** Risk profile methodology; AnteArena internal documentation, available on request: dimension definitions and the limit compliance correction.