Sampling Without Random Access

SteamLens Milestone 2: the sampling study

Five questions: the draw, the size, the error bars, the pollution, the ground truth

Every number verified against its run of record at build time

arda-basarici.github.io

Arda Başarıcı · 2026

The findings, first

SteamLens shows numbers under its claims: “x% of the reviews that mention performance are negative”. Milestone 1 built the census that makes such numbers checkable: 135,260 machine-labeled reviews across 49 games, the labeler’s error measured against human judgment. A live product cannot spend a census on every query. This milestone measures that it never has to:

A fresh game’s report needs at most 2,000 fetched reviews. Pools of 2,000 or fewer are taken whole and quoted exactly; larger pools are sampled at n = 1,000, and every displayed share carries an interval whose 95% promise is itself measured: **0.958 on-corpus calibration coverage, 0.971 on the held-out check** (two fresh games reproduced exactly; one fresh game sampled).

Six terms carry the whole report:

Term	Plain meaning
The census	Every usable English review of the 49 study games, machine-labeled once (Milestone 1). The study’s ground truth: the true share of anything is computable from it.
Coverage	When the report quotes “27% ±3”, how often does the true value actually land inside the bar? The promise is 95 in 100.
Tolerance	How far a displayed share may sit from the true share before it counts as wrong; ruled per share band.
Wilson interval	The standard textbook error bar for a percentage; the simplest formula in the race.
p90	The 90th percentile of errors across draws: the bad-draw tail, not the typical case.
Take-all	A pool of 2,000 reviews or fewer is read whole; the number is exact and no bar is quoted.

The report is one apparatus chapter and five questions. Each question is a self-contained chapter; the details live there, not on this page.

1 · Steam gives no random access. How do we pick reviews? Every implementable draw was raced against true uniform sampling, offline, where the census knows the right answer. Time-proportional windowed won on every slice; Steam’s native newest-first order survives only as a fallback with a disclosed, separately priced bias.

2 · How many reviews do we need? Sample 1,000; read everything when the game has 2,000 or fewer. Validated held-out on three fresh games: the two under the cutoff reproduced exactly (360/360), the one above it sampled at coverage 0.971.

3 · How do we keep the error bars honest? More data quietly makes a textbook bar *less* honest here, because the draw’s bias does not shrink with sample size. The shipped bar adds a measured price for that bias, switched on only for burst-shaped (“spiky”) pools; a 24-game long-tail probe measured overwhelmingly calm, so deployed reports should mostly quote plain Wilson bars.

4 · What happens when the sample is polluted? The certified promise survives 2% review-bomb material and is broken by 5%, and the error bars break before the numbers look wrong. Steam's default bomb-window blanking is thereby certified load-bearing, not cosmetic.

5 · How good is the ground truth itself? Two human instruments bound the machine reference's imperfection: review-level agreement 0.557 [0.477–0.634], reading lowest on the newly trusted material, and 11.6% [6.6–19.6] of displayed claims misattributed, nearly all sibling-label mix-ups. Polarity is near-clean; label ownership at family boundaries is the soft spot.

What this report refuses to claim:

- **Not uniform sampling.** The deviation from uniform is measured and priced, not eliminated.
- **Not spiky-regime transfer.** The spiky allowance is validated on-corpus only; the one held-out spiky game is take-all by construction.
- **Not a resolved 2%-to-5% gap.** No product decision changes inside it, so resolution was deliberately not bought.
- **Not unmarked-bomb safety.** The damage rate is measured; the frequency cannot be.
- **Not reference truth.** The reference's imperfection is bounded, not absent.
- **Not non-English coverage.** One held-out game has 36 usable English reviews of 2,277, quoted as the reality it is.

How to read this report. "The instrument" states the shared apparatus once: the census, the certified population, the determinism that makes replication mean anything. Every number cites its run of record, and the repository regenerates all of them.

The instrument

The census is the ruler

Milestone 1's product is Milestone 2's instrument. The census (135,260 reviews across 49 games, machine-labeled under a frozen triple, the labeler's error measured at F1 0.766 [0.713–0.811] against a human-adjudicated reference) is the ground truth this study samples against: a sampling policy can only be raced against the true uniform answer where that answer exists, and it exists only offline, in the census. The reference is machine-labeled, and the study says so up front: what gets measured is sampling error, while the classifier's own error rides silently on top. Bounding that rider is question 5's whole job.

The certified population: what a claim quantifies over

Windowed draws are fully deterministic. Same corpus, same plan, same sample, and this is true of the live runtime too. A single game queried once is therefore exactly one draw: no repeat variance exists to average over, and a "95% of intervals cover" claim needs a population of report runs to be a statement about.

The certified population is composed of three axes: **query anchors × games × displayed aspects**. A query anchor truncates a game's corpus at fixed quantiles of its own review-time span (40 / 55 / 70 / 85 / 100%, never an absolute calendar grid, which would predate thin-coverage games). Compiling from the truncated histogram reproduces exactly what a live query at that moment would have seen, so the anchors turn one snapshot into a population of plausible report runs, and every ruling in this report is a claim about report runs generally rather than about one date.

Two disclosures ride this construction: anchors within one game are *nested*, widening the population without being independent replications; and truncating today's corpus at time T assumes Steam would have served the same rows then, an approximation only the live tests ground.

Sweep hygiene follows: duplicate truncated pools are dropped, and a cell whose ladder size reaches its pool is recorded take-all and skipped, since a take-all draw's zero error is free flattery for a convergence curve. The one place take-all cells are measured instead of skipped is the closing test, where the cutoff side of the size rule is itself under test.

The two gates

Every certification in this report reads the same pair of gates at the 95% register, over the certified population:

- **Coverage**: the share of cells whose quoted interval contains the census truth. The promise is 95%. An interval method that cannot hold it under the real draw is eliminated or repriced, never excused.
- **Tolerance**: the share of cells whose point error stays inside the band's ruled tolerance. Tolerances condition on census-share band (tail, mid, headline), because a one-point miss means something different at 2% share than at 40%.

Two demotions and one floor complete the grammar. Rank stability of the top aspects and praise/criticism direction are measured and reported but never gate: both follow from shares being right, so gating them would add criteria without adding information. And the display evidence floor, the deployment-layer threshold that greys out thinly-supported aspects (its value is ruled at deployment), can

only narrow what is shown: the study measured every pinned aspect its pools mention, down to zero share, so the promises are floor-independent and the floor never adds an unpromised cell.

The gates are the study's fixed grammar. The curves checkpoint ruled the policies and constants through them, the mixing experiment re-runs them under contamination, and the closing test reads them held-out. One honesty standard runs end to end: the mixing floor is defined as the marked share at which conclusions drift beyond the very tolerance the checkpoint set, not a separately guessed percentage. When a chapter says "the promise holds," it means these two numbers, on this population, at this register.

The runs of record

Every figure in this report regenerates from a named, seeded run; none is hand-carried. The study's spine is four:

- **The curves sweep** `m2sweep-20260802T132010Z-2969bcab`: 49 games, 243 anchor pools, 5,476 cells, 255,744 measurement rows, about five minutes of CPU. The counts nest: a cell is one game × anchor × policy × size draw, each row one displayed aspect's measurement inside it; the rows split by census-share band into 219,296 tail, 30,140 mid, and 6,308 headline, and later slices (a policy, a size tier, a regime) are subsets of these, down to the 48 headline cells the spiky calibration stands on.
- **The long-tail discovery** `longtail-20260802T232206Z-9bf61718`: 959 seeded-uniform probes of a persisted 177,272-app catalogue snapshot, 24 games admitted by criteria alone.
- **The mixing run** `m2mix-20260804T120612Z-c31f92fe`: the certified grid × three bomb sources × an 8-point share grid, 200 seeded blends per cell; at share zero it restates the on-corpus calibration (coverage 0.958, tolerance 0.982–0.983).
- **The closing test** `m2close-20260804T140340Z-1cc06586`: 3 held-out games, 15 game-anchor cells, 605 aspect-level rows (360 take-all reads + 245 sampled cells), the finished rule run as shipped.

The fresh material

The study bought one batch of material the corpus could not supply: six games deliberately disjoint from the 49. Three carry verified review-bomb marks spreading the marked-window population (a canonical tight window, a small ongoing mark, a low-English tight window), each verified on the wire before purchase. Three are long-tail games picked from the discovery run's admitted list to span the regimes it surfaced.

The fresh labels live in their own store inside the fetch run's directory. This is containment by storage: nothing non-census can reach the production pool, so the two-track wall (displayed numbers come from the survey mint only) holds by construction. And because the census's certificate certifies only the census's labels, the buy, the paid labeling run, carried its own: a fresh certification against gold under the frozen triple, F1 0.776 [0.727–0.818] beside July's 0.791 [0.742–0.836]. Two runs of one instrument make a two-point drift series, consistent with the standing buy-time-variance rule from Milestone 1.

The human labels ride the same discipline: the 150-review holdout was drawn seeded, stratified, and deliberately oversampling the fresh strata, labeled blind under the frozen codebook, and journaled as an eval run like any other.

QUESTION 1

Steam gives no random access. How do we pick reviews?

The problem

Every sampling textbook starts from a draw the Steam API does not offer: pick a review uniformly at random. The API serves cursors, newest first, with the date window as the only real steering wheel. So the honest question is which of the draws we *can* implement comes closest, and what the remaining distance costs; both halves need ground truth, which only the census owns. This chapter is the race that fact makes legal.

The race, and why it is nearly free

Four policies ran against each other over the certified population (query anchors × games × displayed aspects, per the instrument chapter), on a size ladder densified at the low end (100 / 250 / 500 / 750 / 1,000 / 1,500 / 2,000 / 3,000 / 5,000). The windowed policies are deterministic, one draw per cell, with the anchor grid supplying the replication; only the uniform reference repeats, at 200 seeded draws per cell. The census pays its dividend here: a simulated draw resamples stored labels, CPU only, zero LLM spend, so density costs minutes rather than dollars. The full sweep, 49 games × anchors × 4 policies × the nine-size ladder (5,476 cells, 255,744 persisted measurement rows), ran in about five minutes of CPU.

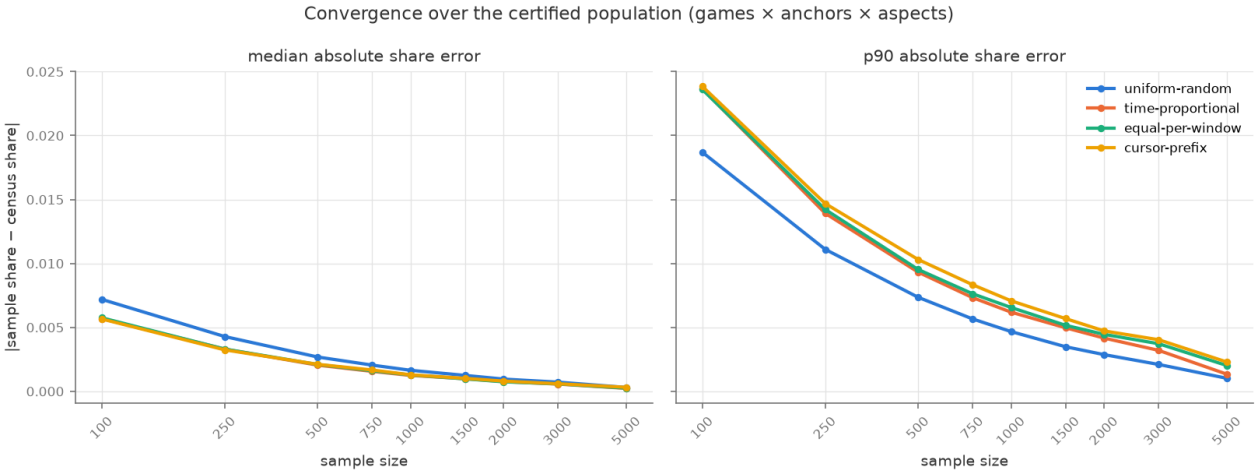
The candidates:

- **Uniform random.** Not runtime-expressible; simulated as the textbook reference every other policy is judged against.
- **Time-proportional windowed.** The runtime primary path's hypothesis: spread the fetch budget across date windows in proportion to each window's review volume, then take each window's newest-first prefix. An approximation of uniform access built from the parts Steam actually offers.
- **Equal-per-window.** The budget spread evenly across windows, over-representing quiet periods by construction; raced so its expected rejection would carry numbers.
- **Cursor-prefix.** The documented fallback as it actually behaves: a plain most-recent prefix, biased by construction; its bias will be quoted to users, and a quoted bias should be a measured one.

Playtime and vote-type balance were deliberately not raced: whether the API can even express them is unverified, and a policy that cannot ship has no business winning a race. They ride as representativeness diagnostics on the winner.

The verdict

The race is invisible at the median: median share error is small everywhere and near-identical across policies, roughly 0.6 points at n = 100 and under 0.1 by n = 2,000. What separates the candidates is the p90, the bad-draw tail, with uniform best and cursor-prefix worst.

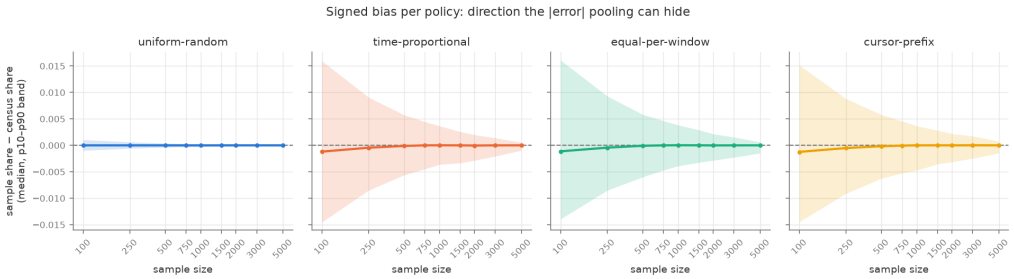


Convergence per policy: median and p90 share error against the census truth, by sample size. Run of record m2sweep-20260802T132010Z-2969bcab.

Time-proportional windowed is the primary path. It led the other implementable draws on pooled p90 share error, per-band error, and Wilson coverage at every n on the ladder. The per-slice margins are often small (only the unreachable uniform reference separates cleanly by eye), but the direction never flips, and the lead is largest in the big-share aspects where a prefix's bias concentrates.

Equal-per-window is eliminated. Its quiet-month over-weighting never paid for itself on any slice: it buys representation for periods with few reviews by misweighting the periods that dominate the true share.

No policy hides a net direction. The sweep's signed-bias view showed symmetric spread for every policy: misses in both directions, nothing to re-center. This mattered twice downstream. It means the point estimate stands as quoted (question 3 inflates only the width, never shifts the center), and it means the fallback's user-facing disclosure is a spread statement, not a drift correction.



The signed-bias view: no policy hides a net direction; the misses are symmetric spread. Same run of record.

The fallback, priced separately

Cursor-prefix survives in its designed role: the path a report falls back to when windowed fetching is unavailable, always disclosed on the trust panel. Its spread at small n is wide (roughly ±1.5 points at p10–p90 for n = 100) but symmetric, and it gets its own price rather than borrowing the primary path's: the fallback calibrates its own allowance constants from the same measurement rows' cursor-prefix column (the constants themselves are question 3's subject). A newest-first prefix needs no temporal spike to be biased, so unlike the primary path its calm regime carries real allowances, and the disclosure quoted to users is regime-aware.

The chapter's claim is deliberately narrow: not that the windowed draw is unbiased (question 3 prices its bias), not that it represents playtime or vote-type strata (diagnostics, not promises), and not that the race generalizes off-population (question 2's closing test is that check).

QUESTION 2

How many reviews do we need?

The problem

The founding plan guessed at this with a ladder of three sizes to try. The census made guessing unnecessary: with ground truth owned, the study can draw at every size on a dense ladder and watch the readings settle against the true value. The deliverable is a rule rather than a number, because corpus games span orders of magnitude in review count and one n cannot be the answer for a game with 900 reviews and a game with 90,000.

The knee in the curves

Convergence, read on the tail and mid share bands where sampling error behaves classically (the big-share bands play by different rules, question 3's subject), settles fast. The median share error is already small at n = 100 and under a tenth of a point by n = 2,000; the p90 follows the expected √n decay. Two facts on the ladder locate the choice:

- **n = 750 passes the ruled tolerance with no margin.** The mid band clears at 2.3 points against a 2.5 tolerance; a rule sized there would bank on off-corpus games behaving no worse than the corpus.
- **n = 1,500 buys 0.2 points of error for 50% more cost.** Returns have flattened; past this the curve pays for fetch and classification without measurably improving what the report displays.

So the sampled size is **n = 1,000**: one tier above the smallest passing size, safety margin bought at the point of maximum marginal return.

The cutoff, and why it is an honesty boundary

The second half of the rule is a population cutoff: **pools of 2,000 or fewer are taken whole.** The shape is 2×n. Below twice the sample size, sampling saves less than half the fetch-and-classify cost, so exactness is nearly free, and the take-all pool quotes its exact number with no sampling interval at all, a census of itself.

The cutoff started the study as a cost knob and ended it as something stronger. Question 3 shows the big-share aspects' error is bias, flat in n, so no affordable sample size fixes them; the only regime that makes a headline number *exact* is take-all. That quietly raised the cutoff's importance: it is the boundary where the product's loudest numbers stop carrying wide intervals and become exact counts.

> Stated as the product will state it: **we read 1,000 reviews; if the game has 2,000 or fewer, we read them all.** Every report costs at most 2,000 fetched-and-classified reviews. That is the founding hypothesis made concrete: not the corpus's three hundred thousand (298,553 raw reviews, 135,260 usable), at most 2,000, with the price of the difference measured instead of assumed.

One honesty mark rides the rule. The study corpus leans long-tail (median full pool about 2,100 reviews), so the fraction of corpus anchor pools that land take-all overstates the rule's real-world reach: the popular games users will actually query sit permanently in the sampled regime, quoting calibrated intervals rather than exact counts. The report says so plainly rather than letting the corpus composition flatter the rule.

Brute force was rejected by the study's own curves. "Just sample more" was a candidate answer at the checkpoint, and the curves killed it: the headline-band error is flat in n, so a larger n buys almost nothing exactly where the pressure is, while cost grows linearly. Raising the take-all cutoff instead survives as the honest lever (exactness is the one thing money reliably buys here), and stays available to the deployment milestone if product needs demand it.

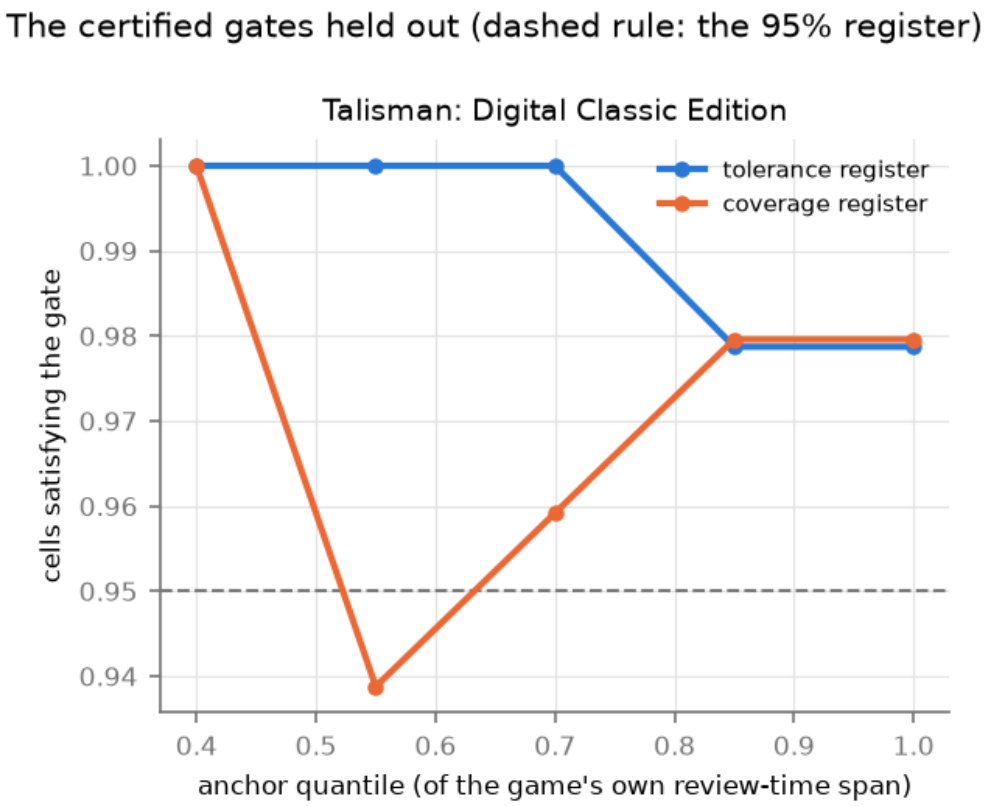
The closing test: the rule meets games it never saw

Every ruling so far was calibrated on the same 49-game corpus it was measured against. The closing test is where the finished rule met three freshly bought games, deliberately disjoint from the corpus, each fully labeled under the frozen triple so its own full-pool fold is ground truth: Sword and Fairy Inn 2 (36 usable English reviews of 2,277, the language-stress case), Dragonkin: The Banished (1,311, the weekly-served young game), and Talisman: Digital Classic Edition (6,094, the flat mid-band case above the cutoff).

Two design rulings shaped the measurement. The full anchor, "a fresh game queried today", is the honest single unit but yields too few cells for a 95% register (deterministic draws mean each cell is exactly one draw), so the certified own-span anchor grid measures, five simulated query moments, and the full anchor headlines. And the spiky-regime conditioning's off-corpus transfer is not a claim here: the lone spiky exemplar is take-all at 36 usable reviews, so nothing sampled exists to validate it; it rides as a disclosure.

One inversion from the sweeps is deliberate: the sweeps *skipped* take-all cells (zero error, free flattery for a convergence curve); the closing test *records* them, exactness-verified, because here the cutoff side of the rule is itself under test.

The verdict: the rule holds held-out. The take-all side delivered exactly what it promises: **360 of 360 recorded reads exact at error zero** across two games. The sampled evidence comes from the one game above the cutoff, five nested anchors: **coverage 0.971 and tolerance 0.991 over 245 cells**, 0.980 / 0.979 at the headline full anchor.



The closing test's register by anchor: both gates against the 95% rule across the five simulated query moments. Run of record m2close-20260804T140340Z-1cc06586. The 0.55-anchor coverage dip (0.938) is the mid-band wrinkle the prose diagnoses; it recovers by the full anchor.

The verdict's honest wrinkle is quoted, not buried: mid-band coverage alone reads 0.902 (46 of 51 cells). Three of the five misses are one aspect (`learning_curve`) at three *nested* anchors, closer to one correlated miss counted three times than to three failures: the nested-anchors caveat with a concrete face. The band read stays diagnosis, not verdict: the certified promise was always the pooled reading, and 51 cells resolve a true 0.95 rate only to about ±3 points of standard error (roughly ±6 at 95%).

The micro-window variant's reopen trigger, "the closing test failing held-out", did not fire.

QUESTION 3

How do we keep the error bars honest?

The problem

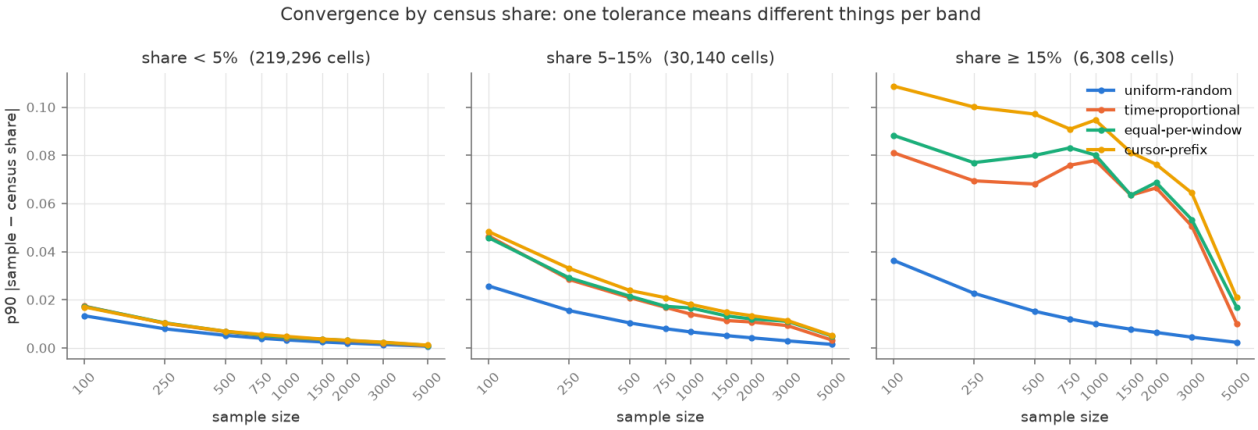
The app never shows a bare 27%; it shows “27% ±3”, and the ±3 is a promise with its own failure mode. Across the certified population the quoted interval must contain the truth at its nominal rate, and this gate can fail while the shares themselves look fine: the numbers right, the confidence around them systematically narrow. For a product whose entire thesis is honest error bars, that failure is the poisonous one. So the interval formula was raced inside the same simulation as the policies, with a pre-committed tiebreak: the *simplest* method whose coverage is honest under the winning policy ships, and the fancier method earns its place only when the simple one’s coverage measurably fails.

The race: the sophisticated method failed its own pretense

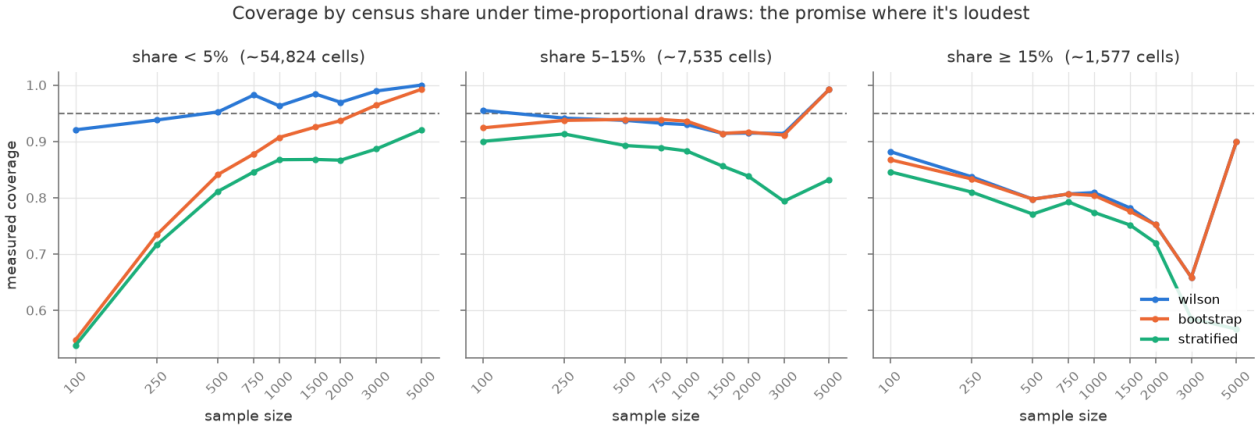
Three candidate formulas were computed on every simulated draw. Wilson, the design-naive binomial interval, held roughly 92–97% measured coverage pooled across all policies and sizes. Bootstrap-over-reviews collapsed at small samples, near 60% coverage at n = 100: with few expected successes per draw the resampled statistic collapses onto a few lattice points and the percentile interval degenerates. The design-aware stratified interval with its finite-population correction, the sophisticated candidate, under-covered persistently (roughly 55–90%): its correction assumes the within-window draw is random when the contract says newest-first prefix. Watching the fancy method fail its own pretense is what the calibration gate was built for; “the simplest honest formula ships” was decided by measurement, not taste. One constraint carried from the eval harness: resampling draws whole reviews, never mentions, since mentions within one review move together and treating them as independent fakes precision.

The finding: more data made the error bars less honest

The pooled verdict, “Wilson covers everywhere”, was hiding something. It was dominated by small-share cells, where absolute errors are tiny by construction. Sliced by the aspect’s census share, the same measurement rows broke the promise exactly where a report is loudest. On aspects at or above 15% census share, the windowed policies’ p90 error sat at 7 to 11 points and barely fell until $n \approx 1{,}500$, because a deterministic newest-first prefix’s error is bias, and bias does not shrink like $\sqrt{n}$; it only dies when growing quotas swallow whole windows. The companion coverage slice confirmed the consequence: on those same aspects every candidate interval under-covered, and coverage got *worse* as n grew, Wilson falling from about 88% at n = 100 to 75–78% by n = 1,500–2,000. The quoted width shrinks like $1/\sqrt{n}$ while the bias stays put, the classic bias-versus-width squeeze. Buying more reviews made the numbers more confident and less honest at the same time.



The p90 error curves by census-share band: sampling is easy exactly where it matters least. Run of record m2sweep-20260802T132010Z-2969bcab.



Measured coverage by share band under the primary policy: on big-share aspects every interval under-covers, worse as n grows. Same run. The n = 3,000–5,000 swings in the headline panel ride thin cell counts; the shipped n = 1,000 does not lean on them.

The ruling: price the pretense

Four candidate answers came to the checkpoint, and each met a different fate. Brute force (larger n, earlier take-all) was rejected by the curves: headline error is flat in n while cost grows linearly. Band-aware tolerance proved necessary but insufficient: re-labeling the tolerance leaves an interval that claims 95% and delivers 75. The micro-window variant, the one candidate that attacks the *cause* rather than repricing the symptom, was parked rather than killed, carrying an unsolved compiler question and an unknown payoff without a re-sweep; it waits on named triggers (the closing test failing held-out, which did not happen, or deployment finding the headline widths product-unacceptable).

What won is the bias-aware interval, and the reason is the study's own thesis applied once more: the whole race had been about pricing pretenses (Wilson won it because its pretense was least wrong), so the consistent move is to pay for the windowed pretense explicitly. The shipped interval quotes Wilson's width plus a measured per-band constant allowance. The signed-bias view (question 1) made the shape clean: no policy hides a net direction, so there is nothing to re-center; the point estimate stands and only the width inflates. Take-all pools quote the exact number and no interval at all.

The checkpoint's constants, minted live from the run of record: 0.000 for tail aspects (Wilson alone covers), 0.005 for mid, 0.073 for headline, each the maximum of the calibration over the shipped tier and its neighbors, deliberate conservatism against order-statistic noise in a thin band. A headline aspect would

ship at roughly ±10 points. The tolerance table landed beside it: ±1 point for tail aspects, ±2.5 for mid, and headline aspects carry no separate error tolerance, because their promise *is* the calibrated interval plus take-all exactness; a tolerance number there would either restate the width or claim precision the draw cannot deliver.

The refinement: the flat constant was an average of two games

The checkpoint ruling was one day old when the within-corpus shape splits indicted it. In the first cut all three shape axes looked guilty (the spikiest tercile at p95 error 3.5 points against ±2.5, coverage down to 0.872), but conditioning untangled it: with the spiky third of anchor pools set aside, the pool-size effect vanishes entirely, and a near-uniform cross-tab showed the axes were all proxies for spikiness. The whole windowed penalty lives in pools whose busiest window holds a large share of all reviews, the same mechanism the policy race exposed, now located instead of averaged.

Located, it showed the flat constants to be an average of two very different games: calm pools that need no allowance at all, whose measured coverage rides near 100% and which had been shipping ±10-point headline bars where ±2.5 suffices, and spiky pools that need roughly double the flat price. That is the checkpoint's own failure mode one level down: over-cautious where it is easy, over-confident where it is hard. The ruling followed the thesis a third time: condition the price on the regime.

Two facts made the conditioning cheap. Peak window share is computable from the live review histogram *before any draw*; the runtime fetches that histogram to plan windows anyway, so the regime adds no data dependency. And the threshold barely mattered where it could have hurt: over a sweep of candidate cuts from 0.50 to 0.75, the calm side minted 0.000 at every cut, so only the spiky calibration hinged on the choice. Two-thirds won over one-half because the lower cut dilutes the spiky pool with borderline units that need nothing, under-protecting the genuinely spiky tail (an 0.109 calibration against the 0.127 those units actually measure).

	Tail (<5%)	Mid (5–15%)	Headline (≥15%)
Primary, calm	0.000	0.000	0.000
Primary, spiky	0.000	0.017	0.127 (~±15 pts)
Fallback, calm	0.000	0.004	0.065
Fallback, spiky	0.000	0.022	0.130

The shipped constants supersede the flat ones, and one caveat meets the reader here rather than only in the limits: the spiky calibration rests on 48 on-corpus headline cells, and no off-corpus sampled draw has ever exercised it. The fallback's calm-regime allowances are themselves a finding: a newest-first walk over the whole pool is biased with or without a spike. A companion ruling closed a tolerance gap: spiky mid joins the headline treatment (no separate error tolerance; ±2.5 is unmeetable there regardless of the interval quoted), while calm mid keeps its ±2.5. The tolerance table is regime-aware in exactly one cell.

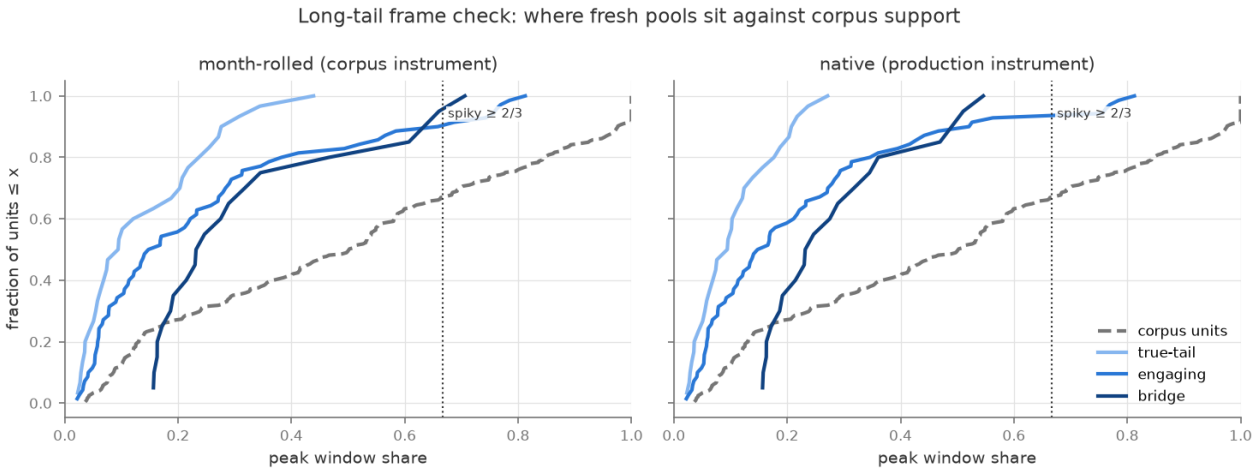
One reproducibility note: graduating the allowance computation from session scratch to a committed mint script surfaced a near-miss reconstruction (0.004 / 0.067), chased until the ruled constants reproduced exactly under the centered-plus-ceiling definition, which errs conservative.

Off-corpus: which regime is the real world in?

The conditioning raised the stakes on a question the study was always going to ask: the corpus is 49 popular games in a recent window, and the deployed app will be pointed at anything. If the long tail were disproportionately spiky, the ±15-point regime would be its normal and the calm constants a popular-game privilege.

The game list that answered this was built so nobody picked it: three review-count bands aligned to the ruled cutoff, filled by seeded uniform probing of a persisted 177,272-app catalogue snapshot, admitting a candidate exactly when the store called it a game and its totals landed in an open band. 959 probes, 24 games, the list re-drawable from the recorded seed and snapshot.

The answer, scoped to its sample: **the long tail measured calm in these 24 criteria-drawn games.** One game in 24 sits in the spiky regime on the production instrument; at the replication grain, 5 of 120 (game, anchor) units, 4.2% against the corpus's 33.1%. Fresh peak window shares (0.022–0.813) sit entirely inside corpus support (0.036–1.000), so the conditioning never extrapolates. Deployed against the long tail, the runtime will overwhelmingly quote the calm constants, plain Wilson bars, with the spiky treatment reserved for the rare game whose whole life is one event.



The frame-check ECDF: fresh long-tail peak shares against the corpus, the 2/3 regime line marked. Run of record longtail-20260802T232206Z-9bf61718.

The comparison also reframed the corpus's 33% spiky rate: the same games' whole-life histograms read far flatter than their corpus-window readings (0.503 falling to 0.042 on the clearest row), so the rate was largely a property of *windowed pools*, not of popular games. The calibration stands, because the spiky constants priced the penalty mechanism, one window swallowing the draw's quota, and the mechanism transfers by shape rather than by span; production, reading whole-life histograms, simply meets it more rarely. Two instrument disclosures ride the evidence: the regime is computed on Steam's native rollup buckets (weekly for 5 of the 24 admits), the shape the windowed compiler actually plans over, with no admitted game flipping across the 2/3 boundary by unit choice; and fresh whole-life pools exceed corpus pool support on the high side (to 63,000 against 6,900), disclosed rather than conditioned on, pool size being the axis the splits cleared.

The chapter's transfer claim is split honestly: calm-regime transfer on held-out evidence, spiky-regime transfer on mechanism only, with the spiky calibration resting on 48 on-corpus headline cells.

QUESTION 4

What happens when the sample is polluted?

The problem

Everything so far assumed the pool is what it claims to be: players reviewing a game. Steam's known failure of that assumption is the review bomb, a coordinated burst of reviews about something else (a pricing decision, an exclusivity deal, a patch), and Steam itself marks such windows and blanks them from default listings. Since Milestone 1 the system had carried a provisional guess that past some share of bomb material a report degrades dishonestly. This experiment replaces the guess with a measured floor: the last marked share at which the certified promise still holds.

The corpus holds zero marked-window reviews, so this was the study's one question that had to run on bought material: 6,445 labeled marked-window reviews from three games with probe-verified bomb marks spreading the population (Borderlands 2's canonical two-week Epic-exclusivity bomb, Book of Demons' small ongoing regional-pricing mark, The Witcher 3's low-English tight window). The wire probe corrected the web research twice before purchase.

The design: one honesty standard end to end

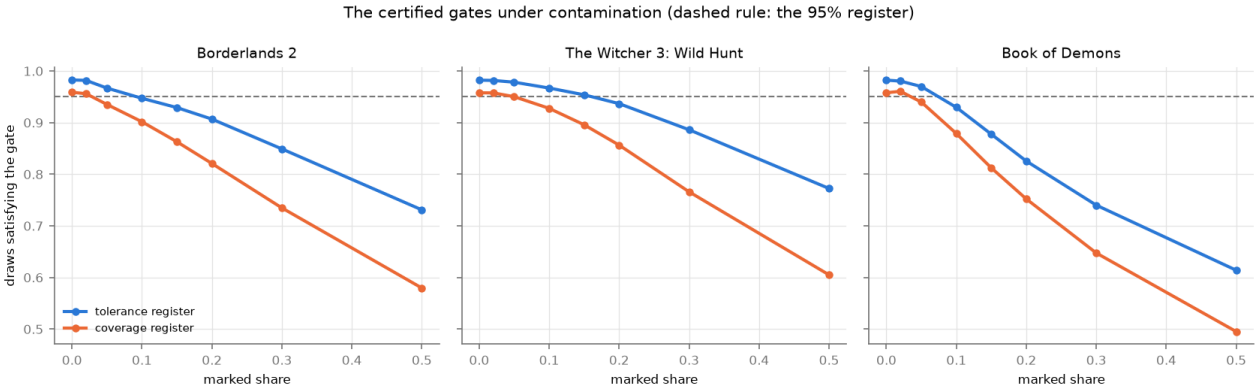
The design had one genuine fork: what does a contaminated number measure against? Measuring against the unmixed sample's own conclusion isolates the marginal contamination effect, cleaner as a pure measurement, but it applies the certified tolerance to a quantity the tolerance was never minted for. The ruling went the other way: the drifted number measures against the census share, the study's exact gates re-run with contamination, so the floor means "the last share at which the certified 95%-register promise still holds". One honesty standard runs end to end, on the same pass/fail machinery production is certified by.

The rest followed. Contamination is replacement at fixed n (a polluted report run is a same-size sample with a fraction being bomb material; adding would entangle a size effect), the share grid densifies at the low end (0 / 2 / 5 / 10 / 15 / 20 / 30 / 50%), and three per-source curves stay separate with the floor read from the worst, since the bombs were picked to differ and pooling would average that away. The base cells reuse the certified grid, 200 seeded blends per cell, entirely offline.

One quiet design decision earned its keep in the first smoke: bomb material *invents* aspects the base game barely has (Borderlands 2's marked window is 25% `platform_access`), so measurement runs over the union of the base and bomb vocabularies, giving such aspects a true near-zero reference. In the smoke, an aspect at 0.36% of the base game inflated to 12.4% of the sample at half contamination, a fabricated headline a base-vocabulary measurement would never have looked at.

The verdict: the floor is 2%

The share-0 baselines pass across all three sources (coverage 0.958–0.959, tolerance 0.982–0.983), which means the run restates the checkpoint's on-corpus calibration before any contamination and the floor is measured against a verified control. Per-source floors: Borderlands 2 0.02, Book of Demons 0.02, The Witcher 3 0.05. The worst source rules: **the marked-share floor is 2%.** The break is grid-located, the promise holding at 2% and broken by 5%, and resolution inside that interval was deliberately not bought, because no product decision changes with it.



Both gates against marked share, per bomb source, against the 95% rule. Run of record m2mix-20260804T120612Z-c31f92fe.

The mechanism is the finding to lead with. **Coverage, not share error, is the binding gate everywhere.** Wilson's width depends on the sample size, not on what contaminated the sample, so contamination shifts the displayed numbers while the error bars stay exactly as confident as before. Headline-band coverage falls to 0.93 at 5% contamination and 0.78 at 10%, long before the raw errors grow conspicuous. The error bars fail silently first: the numbers still look right while the bars around them have stopped being true.

The product meaning

Steam's default listings blank marked windows, production's fetch inherits that default, and the wire probes verified the blanking on these actual games. A production sample therefore carries ~0% marked material by construction, the passing column of every verdict table. What the floor changes is the exclusion's status: from an inherited default to a certified load-bearing requirement with a number attached, since even a 5% admixture voids the calibrated bars. Marked windows stay display-only episode markers on the timeline and are never folded into displayed numbers.

The named residual: an *unmarked* bomb bypasses the blanking and lands in samples as ordinary reviews. This experiment measures that scenario's damage rate, not its frequency, which is unmeasurable by construction; a bomb nobody marked is a bomb no query can count. No usable frequency proxy was found either: temporal spikiness measures burst shape, not intent, and ordinary release and update bursts dominate it, so it bounds burstiness, never bombs. Partial mitigations exist (the spiky allowance prices burst-shaped pools, and the timeline's episode markers make anomalous windows visible to the reader), and the limitations section names the residual plainly.

Two verdict-hygiene notes survive in one sentence each. The first full sweep was killed and re-fired when the analyzer's design showed its summary rows could not reconstruct per-draw pass rates: decide what question an artifact must answer before producing it. And the one-game smoke read the 5% share three points more flattering than the full 49-game grid (0.958 against 0.940), which is why the floor quotes only off the full grid.

QUESTION 5

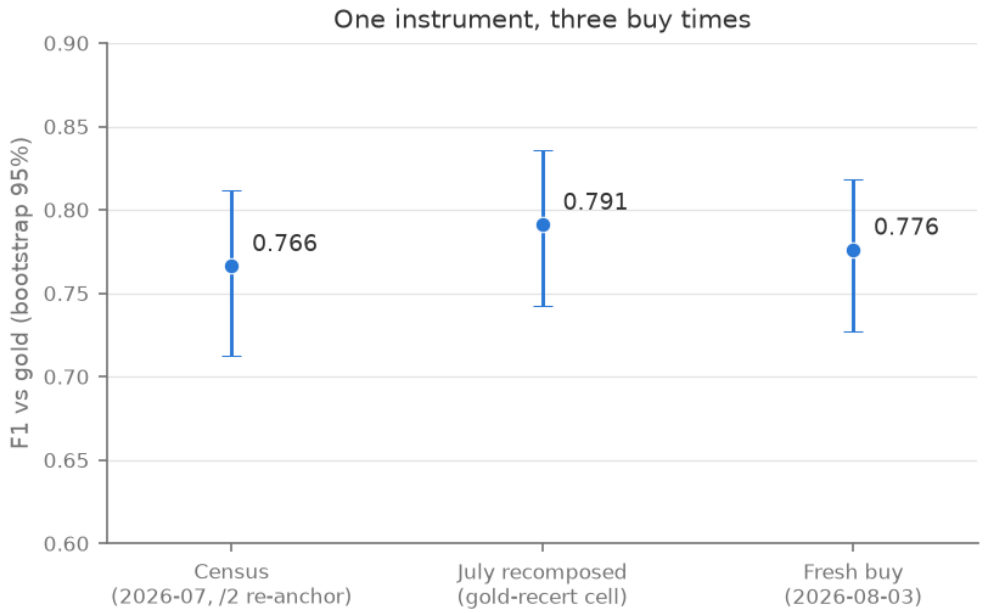
How good is the ground truth itself?

The problem

Every number so far is measured against the census fold, and the census is machine-labeled: the study measured sampling error while the classifier's own error rode silently on top, in territory (marked windows, long-tail games) that Milestone 1's gold set never covered. Two human instruments price that lean: a holdout asking how far production sits from a careful human read, and an audit asking whether the model's verbatim quotes are attached to the right claims. They measure different things, and they converged on one diagnosis. They also mark the edge of a named assumption: questions 1 through 4 treat sampling error and labeling error as separable, and whether the classifier reads burst-shaped or bomb-adjacent material worse is an open question the stratum gradient below begins to probe.

First, instrument identity: the annotator under the fresh labels

The fresh material was labeled months after the census, and the standing rule from Milestone 1 is that a re-buy is never trusted on an old certificate (the served model wobbles at buy time even at temperature zero). So the buy carried its own certificate: a fresh certification against gold under the frozen triple, F1 0.776 [0.727–0.818], sitting between the census's 0.766 and July's recomposed 0.791. A three-point buy-time series, entirely inside the wobble the Milestone 1 experiments measured: the annotator under the fresh labels is the same instrument the census certified.



One instrument, three buy times: the certificate series against gold, each with its bootstrap interval. Journaled certify runs, scorer /2 era.

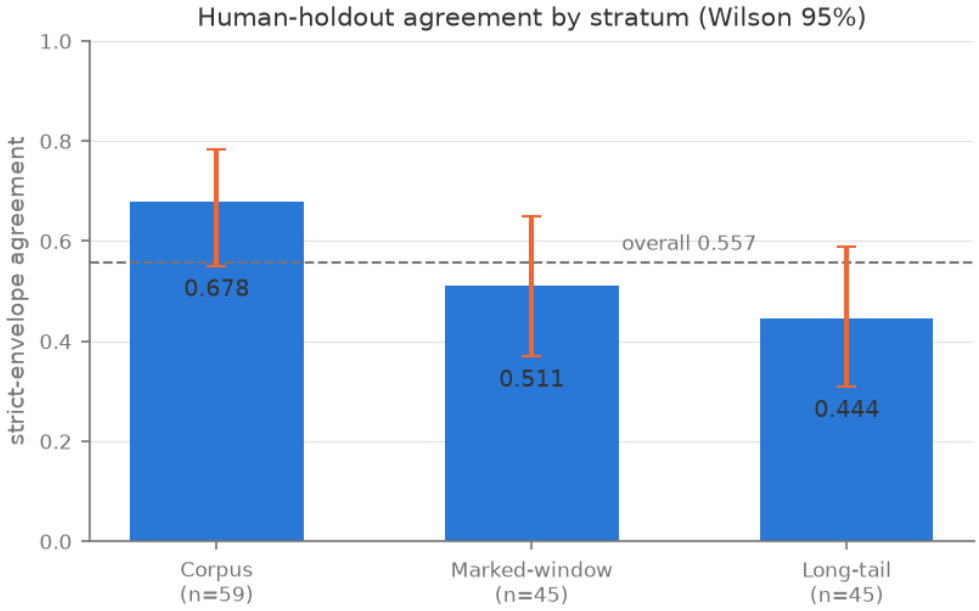
One methods lesson rides the certificate: the first attempt reused July's composition seed and the content-keyed spend cache replayed July's responses at $0.00, so the cell re-ran on a fresh seed. Spend-safety caching and drift measurement are structurally opposed; a drift instrument must vary its content, or the cache will faithfully hand back the past.

The holdout: 150 reviews, labeled blind, drawn where trust is newest

The holdout draw deliberately oversampled the fresh strata: 60 corpus / 45 marked-window / 45 long-tail, seeded, rendered blind (strata and games held back in the machine record only), labeled under the frozen codebook v2 by the project's human annotator, and scored strict-envelope against production labels: a review counts as agreement only when production's pinned aspect set *and* every matched sentiment equal the human's.

The headline is **0.557 [0.477–0.634]** over 149 scored reviews (one non-English skip). Two things must be said next to that number or it misleads. It is the harshest honest bound: one extra or missing aspect fails the whole review, and a ten-mention review needs all ten matches to score one agreement, so it does not share a ruler with the mention-level certification F1 of 0.766. And its decomposition changes what it means: **sentiment-given-matched-aspects is 0.988** (83 of the 84 aspect-matched reviews). When the two readers agree on what a review discusses, they almost never disagree on how the reviewer feels about it. The entire disagreement is aspect selection.

The stratum gradient is the study-design payoff, and the limitations section's spine: corpus 0.678, marked-window 0.511, long-tail 0.444, each at roughly ±12 points. At n = 45–59 per stratum the intervals overlap, so the ordering is directional rather than sharply resolved, but agreement reads lowest exactly where the draw was aimed: **the reference looks weakest exactly where the study newly trusts it.** The sampling promises of questions 1 through 4 are intact as statements about sampling; what this bounds is how far the sampled-and-classified number can sit from a careful human read of the same reviews.



The three-stratum agreement gradient with Wilson bars: the limitations story in one chart. Run of record holdout-20260804T215600Z-c0edb01a.

The pass ran in review batches of ten with production labels walled off, under a drift guard: a mid-pass labeling ruling is checked against gold's recorded applications before adoption, since new territory rules freely but contradicting demonstrated precedent is drift. The guard's first live use flipped the question that prompted it, and it cut both ways between the two readers. One hygiene incident earned a permanent rule: a format-on-save pass silently stripped whitespace inside the sheet's review fences (21 of 150 drifted from the machine record); the sheet was restored byte-true from git history, and every human labeling surface is

now formatter-ignored, because an evaluation certifying verbatim quotes must keep its labeling surfaces byte-true.

The audit: are the quotes attached to the right claims?

The misattribution audit reads 100 displayed claims, each a verbatim quote the fabrication gate already verified, and asks whether the quote supports the aspect and sentiment it is filed under, judged in the review's context (judged as bare spans, the rate would have measured how tersely the model quotes; the frame ruling flipped what the number means).

The verdict: **11.6% [6.6–19.6]** of decidable claims misattributed (95 decidable of 100; two non-English primaries forced a disclosed skip-and-replace under a coverage-enforcing reserve). The decomposition matches the holdout from the other side: aspect-side failures 10.4%, sentiment-side 3.1% (two claims fail both sides, so the sub-rates overlap rather than sum to the total). And the failure profile is the interesting part. Nearly every aspect miss is close-family routing: crashes filed under `bugs` where the codebook puts hard failures under `stability`, developer-incompetence rants filed as `updates`, enemy stat-tuning filed as `ai_behavior` where the codebook routes number-blame to `balance`, plus a small class of wish-quotes (feature requests read as evaluations, where the codebook rules an absent feature is not an aspect of the game). Zero far-field misreads: no music quote labeled as graphics anywhere in the sample.

One verdict from two instruments

The model reads polarity reliably and fumbles label ownership. Sentiment survives both instruments near-clean (0.988 given matched aspects; 3.1% sentiment-side misattribution). What fails is which sibling label owns a mention, at family boundaries the codebook itself had to draw lines through. For the displayed numbers this locates the reference's imperfection precisely: a share for `bugs` and a share for `stability` each carry boundary noise that their sum largely does not, and the praise/criticism direction of any aspect is trustworthy well past the aspect split itself.

Agreement is not accuracy: adjudicating who is right where the readers part remains the parallel human track's job. And one held-out game is not humanly checked at all: the language-stress case drew zero holdout reviews under the uniform within-stratum draw (expected 0.9).

CLOSING

What this earned, and what it didn’t

What the milestone set out to test

Milestone 1 closed with a promise: its numbers rode a fixed census, and the sampling study would build the estimator discipline for defensible numbers from live sampling, where no census exists and the API offers no uniform draw. The founding plan guessed at this layer (a ladder of three sizes to try, a provisional bomb-share threshold); the study’s job was to replace every guess with a measurement or with a named refusal to claim.

It did. The deliverables are a size rule validated on games it never trained on, an interval whose 95% promise is measured rather than assumed, a contamination floor locating where that promise breaks, and a measured bound on the reference’s own imperfection. Two of Milestone 1’s own asterisks resolved inside this milestone: the registered fresh holdout ran (its number is question 5’s spine), and the buy-time variance rule was exercised for real, twice, on the fresh buy’s certificate.

Limits, plainly

- **Reviews, not players.** Unchanged from Milestone 1 and always true: every number is a share of reviews in the labeled pool. Reviewers self-select; no claim about the player base is made anywhere.
- **A popular-game corpus calibrated the constants.** All allowance constants are self-calibrated on the 49-game corpus. The calm-regime promise passed a held-out test; the spiky-regime allowance has no off-corpus sampled check (the one held-out spiky game is take-all by construction) and rests on 48 headline cells even on-corpus.
- **The certified population is a construction.** Query anchors within one game are nested, not independent replications, and truncating today’s corpus at time T assumes Steam would have served the same rows then; edits and deletions make that an approximation only the live tests ground. Register reads are cell-weighted and cells share games and anchors, so effective replication is smaller than cell counts suggest; a game-clustered reading is a named improvement, not yet run.
- **The reference is machine-made, and its imperfection is now a number.** Strict-envelope agreement with a careful human read is 0.557 [0.477–0.634], weakest on the material the study newly trusts (marked-window 0.511, long-tail 0.444), with aspect selection accounting for nearly all of it and 11.6% [6.6–19.6] of displayed claims misattributed, close-family routing dominating. Sentiment is near-clean on both instruments.
- **Agreement is not accuracy.** Which reader errs where they part is the parallel adjudication pass’s job, still pending.
- **The unmarked bomb is unpriced by construction.** The floor measures what marked-grade contamination does to a sample; a bomb nobody marked bypasses the blanking and cannot be counted by any query. Partial mitigations: the spiky allowance prices burst-shaped pools, and the timeline’s episode markers keep anomalous windows visible.
- **The 2%-vs-5% gap is unresolved on purpose.** The floor is grid-located; no product decision changes inside the gap, so resolution was not bought.
- **English-only, still.** The instrument reads English; for one held-out game that means 36 usable reviews of 2,277, quoted as the production reality it is.

- **Nothing is deployed yet.** The rule, constants, and floors exist as versioned artifacts and runs of record, not behind a URL.

The road

Deployment is next. The size rule ships as stated; Steam’s marked-window blanking ships as a certified requirement rather than an inherited default; marked windows appear only as display-only episode markers on the timeline. Two triggers from this study ride along: if deployment finds the spiky regime’s headline widths product-unacceptable, the parked micro-window variant reopens, and the deployed host re-verifies reachability and rate budget from its own network before anything else. After deployment, the interrogation chat, whose evaluation is already runnable offline against the census.

> ***The pretense this study refused was uniform access; what it built instead is a sampling layer that knows its own distance from uniform. At most 2,000 reviews per report, with the price of every shortcut measured: that is what “you don’t need three hundred thousand reviews” means when it is earned rather than hoped.***