The Goodhart Horizon: A Self-Monitoring Impossibility Theorem for Proxy Optimization, and the Price of Seeing Past It

Ildar Fazulyanov · Independent Researcher, IQ Valuations · dar@iqvaluations.com

Preliminary working paper · v1.8 · July 2026.

Status note. This version combines the readable working paper with the proof kernel in technical appendices. The core closed forms reproduce under symbolic and high-precision checks, but the mathematical proof package has not been human-refereed. In particular, the realized-tilt product-form budget-insufficiency certificate in Section 4 and Appendix C should be treated as provisional until a human mathematical referee signs off; prospective pricing of a not-yet-realized deeper tilt remains open.

Abstract

Optimizing against a proxy eventually diverges from the welfare the proxy was meant to track. This is Goodhart's law, and in reinforcement learning from human feedback and reward-model overoptimization it is visible as a curve: true reward rises, peaks, and falls while the proxy keeps climbing. This paper shows that relative to any monitoring channel the system can build from its own trajectory, the divergence has a horizon, a structural limit beyond which the system cannot see whether it has crossed. We call it the Goodhart Horizon, and establish three results about it. Each result is conditional on a stated premise about what the monitor can read, and exactly as strong as that premise. First, the Horizon. The decisive parameter, the Goodhart gap (how much welfare is given back past the peak), is non-identified from any monitoring channel whose conditional law given the system's trajectory does not depend on welfare, the class that is closed under functions of the system's own trajectory, including, when the welfare deficit is outcome-defined rather than mechanism-encoded, transcript-graded evaluation and trajectory-based interpretability. Two welfare worlds, one benign and one collapsing, generate identical observation laws; no test, estimator, eval, or stopping rule computed from any such channel separates them above chance, at any sample size. The impossibility is structural, the decision-relevant quantity leaves no signature in any channel the system can build from its own trajectory, a blindness we borrow the rhetorical shape of self-reference limits to name, claiming no Gödel-style diagonalization or fixed-point construction. Second, the price of seeing past it. Escape is architectural: it requires a channel whose readings depend on welfare itself, and that channel is priced by a closed-form overlap functional whose growth regime (polynomial, stretched-exponential, or Gaussian-exponential) is now fixed by a regime-spectrum lemma in terms of the thinness of the proxy's upper edge (with the dramatic regimes an edge asymptotic, not instantiated by the measured instance below). Third, the audit hierarchy. Price visibility is structurally one-sided: internal telemetry can expose undercoverage, while sufficiency requires a welfare-touching channel or an explicit worst-case allowance; certifying any grader's informativeness is itself welfare-touching, so each level of oversight requires a new welfare-touching certifier one level up. A governance reading follows: an institution that tries to audit itself reproduces the Horizon one level up, so a stable institution must import exogenous outcome-contact, a welfare-touching channel it cannot compute from its own trajectory, of which mandatory external audit is one realization, not the requirement itself; a companion in preparation argues that, under stated substrate-dependent conditions, no mechanism jointly satisfies incentive compatibility, participation, budget balance, share-of-gains, and a worst-case alignment ceiling, summarized there, not proven here.

Keywords: Goodhart's law, proxy optimization, self-monitoring, non-identifiability, impossibility theorem, importance sampling, minimax lower bounds, reward hacking, AI oversight pricing.

1. Introduction

In 2023, Gao, Schulman, and Hilton measured what happens when a policy is optimized ever harder against a learned reward model. Gold-standard reward rises, peaks, and falls while the proxy climbs throughout. To draw that curve they needed a second, gold reward model, gold by stipulation, a channel whose readings depended on welfare by construction. Khalaf et al. (2025) proved the single-peaked shape for tilted families under total

positivity, and their threshold-finder requires samples of the true reward. Karwowski et al. (2024) derived a worst-case stopping rule for Goodhart effects in Markov decision processes, supplied with a bound imported from outside the training signal.

Every method that locates the peak imports information from beyond the optimizer's own data. This paper proves that the pattern is a theorem (there is a horizon past which the system's own channels cannot see) and then prices the import needed to see past it. The limit is structural, fixed by what the system can measure about itself, and is in the lineage of self-reference limits where a system cannot establish from inside what is only visible from outside (Gödel 1931; Tarski 1936); we borrow the shape, not the machinery, and claim no diagonalization or fixed-point construction. Loosely, the trajectory closure plays the role of the formal system and W a quantity it cannot express; the analogy breaks in that there is no self-reference and no enumeration of statements. The mechanism here is ordinary statistical non-identifiability.

The contributions are three. (i) A self-monitoring impossibility theorem, the Goodhart Horizon: the Goodhart gap is invisible to every monitoring channel whose law is fixed given the trajectory, with the exact boundary of that impossibility drawn. (ii) A pricing suite: escape costs are cheap looking backward, governed by a closed-form overlap functional looking forward, with the growth regime (polynomial through Gaussian-exponential) now fixed by a numbered regime-spectrum lemma in the thinness of the proxy's upper edge. (iii) The audit hierarchy: the price is estimable from internal data only in the direction that says "not enough," and every level of oversight that would certify the next is itself welfare-touching.

Two scope statements up front. The single-peaked shape is classical, following from Karlin's variation-diminishing theorem under total positivity; I claim no priority on it. And the impossibility is conditional on a channel premise, stated below as a premise and defended, not smuggled.

2. Setup

A probability space carries a bounded proxy u and a bounded welfare functional W, with $|W| \leq B$. Formally: probability space $(\mathcal{A}, \mathcal{F}, \mu_0)$ with proxy u and welfare W both bounded and measurable; $\nu = u\_\mu_0$ is the pushforward proxy law; $\pi\_\Phi \propto e^{\Phi u}\mu_0$. Write ν for the law of the proxy, with finite upper edge s (its essential supremum), and $w(s) = E[W \mid u = s]$ for the conditional welfare mean. Optimization is exponential tilting: at pressure $\Phi \geq 0$ the policy is $\pi\_\Phi \propto e^{\Phi u}$, which is exactly the KL-regularized reinforcement-learning optimum (Korbak et al. 2022), so Φ indexes how hard the system is optimizing. Φ is the inverse regularization weight, not a KL budget; the realized divergence $KL(\pi\_\Phi \| \mu_0)$ increases with Φ. The welfare trajectory is $T\_w(\Phi) = E\_\{\pi\_\Phi\}[W]$.

Two identities make the rest go. The trajectory's derivative is a covariance, $T\_w'(\Phi) = Cov\_\{\pi\_\Phi\}(u, W)$: welfare rises under pressure exactly while proxy and welfare remain positively correlated under the current tilted measure. And the trajectory depends on the welfare functional only through the pair (ν, w). Goodhart's law is then the statement that optimization changes the measure until the covariance turns. To avoid assuming an ordinary tail limit that need not exist, define the tail floor $L\_- = \liminf\_\{\Phi\to\infty\} T\_w(\Phi)$ and the Goodhart gap $\Delta = \sup\_\Phi T\_w(\Phi) - L\_-$. When the ordinary limit $L = \lim\_\{\Phi\to\infty\} T\_w(\Phi)$ exists, this reduces to the earlier expression $\Delta = \sup\_\Phi T\_w(\Phi) - L$.

The channel premise (OC). A monitor observes a process whose conditional law given the action trajectory does not depend on W. This covers every statistic of realized payoffs, reward-model scores, transcripts, and policy weights or activations, because the trajectory's law is a function of (proxy, base measure) alone. It also covers any grader that is itself a fixed function of the trajectory, including transcript-graded evaluation when the welfare deficit is outcome-defined rather than legible in the transcript. It does not automatically cover live labels, red-team interventions, deployed A/B outcomes, teacher signals, incident reports, or settlement data; those channels must be classified by whether their output law would change across two deployments with identical transcripts but different downstream welfare. It fails exactly for channels whose output law depends on W: gold welfare measurements, outcome audits, interventions, and graders that perceive the deficit. A practical test: ask whether two deployments with identical transcripts but different downstream welfare would change the channel's output distribution; if not, the channel is (OC). (OC) is a maintained causal premise, not something testable from the data it governs. The impossibility below is exactly as strong as (OC), and no stronger.

Lemma 1 (trajectory closure). Anything the system can compute about itself is a measurable function of its action trajectory, whose law is the (μ_0, u)-pushforward; the welfare functional W enters none of it.

Proof. The trajectory law is the pushforward of μ_0 under the policy, a function of (μ_0, u); any internal reading is a measurable function of the trajectory, so its law is determined by that pushforward. ∎

Combined with the channel premise (OC), that the monitoring kernel is fixed across welfare functionals, the joint law of trajectory and reading is invariant to W. That W-invariance is not derived here; it is the content of (OC). This is the self-reference structure of everything that follows: the system's knowledge of itself is closed under functions of its own trajectory, and welfare lives outside that closure, but the claim that no internal reading moves with W is the premise, not a theorem. The Horizon is the boundary of that closure.

3. The Horizon

Theorem 1 (the Goodhart Horizon). Fix the proxy and base measure and any channel satisfying (OC). The joint law of trajectory and monitor output is a function of the trajectory law and the fixed channel kernel, and is invariant to W (Lemma 1 under (OC)). Consequently all welfare functionals generate identical observation laws. On nondegenerate proxy laws with arbitrarily small baseline mass in top bands, worlds with $\Delta = 0$ and worlds with Δ growing toward the bound (always $\Delta \leq 2B$) are observationally identical. No test, estimator, eval, or stopping rule computed from (OC) data distinguishes such worlds above chance, at any sample size.

Proof sketch. Take two welfare functionals that agree away from the proxy optimum and differ only by a nondecreasing "ramp" supported on a thin top band near s*. The two trajectories differ by the integral of the ramp against the tilted measure, which tends to the ramp height as pressure grows and is at most the top-band baseline mass at zero pressure. So when the top band carries little baseline mass, the gap differs between the worlds by an amount that grows with the ramp height, up to the welfare bound. But neither the trajectory law nor the channel kernel changed, so the observation law is identical, and identical laws give Bayes error one-half against the two-point prior. The full construction realizes both worlds with W deterministic given u, which pins the off-band conditional laws that a mean-only specification would leave free.

Two honest scope notes. What is shown non-identified is the gap Δ; the exhibited pair shares the same boundary peak location, so non-identification of a distinct interior peak is a separate question not settled here. And the large-gap clause needs the top band to carry small baseline mass, which the construction states explicitly.

The Horizon is the geometry of what the channel can express, not an adversary hiding something and not a noise floor. The information was never in the readings. When misalignment is instead encoded in the mechanism itself, deception or sandbagging, the construction does not apply in principle and the theorem is silent; interpretability targets exactly that complementary regime.

Scope of the interpretability implication. The two worlds differ only in W, a latent outcome function with no causal path into the system, so weights and activations are identical and uninformative in principle, but only in the regime where the welfare deficit is outcome-defined (code that reads well versus code that runs; engagement versus downstream harm). When misalignment is encoded in the mechanism itself (deception, sandbagging), the construction does not apply and the theorem is silent; interpretability targets exactly that complementary regime.

Remark (two escapes from the Horizon). Two escapes from (OC) exist: channels whose readings depend on welfare itself (gold evaluations, outcome audits, ρ-graders) and channels that intervene to change the regime under which data is generated (off-distribution red-teams, held-out behavioral evals, deployment-time interventions, the Lucas-critique pattern that regime violations are detectable only through regime variation). Both step outside the closure of Lemma 1; that is what "seeing past the Horizon" means.

Ancestors and adjacency. The closest formal relative is Lang, Foote, Russell, Dragan, Jenner, and Emmons (2024, NeurIPS): they characterize an irreducible ambiguity in the return function for a specific observation channel and learning rule; the theorem above is channel-universal over (OC), and §4 prices the escape. Value unidentifiability from behavior is classical (Ng-Russell 2000; Armstrong-Mindermann 2018), and its POMDP form (distinct latent environments behaviorally indistinguishable) is Armstrong (2018); the new content here is that the operationally decision-relevant gap Δ inherits exact non-identifiability under every fixed-kernel channel (including

the optimizer's own internals) and that §4 prices the escape one-sidedly. The ELK problem (Christiano-Cotra-Xu 2021) is the engineering form of the same boundary; Manheim-Garrabrant (2018) taxonomize the phenomenon Δ quantifies. This is the missing-overlap engine of partial identification (Khan-Saveski-Ugander 2024; Manski 2003), but Δ is non-identified under a maintained channel premise rather than bounded by a smoothness restriction, and §4 prices buying the overlap back rather than bounding it.

4. The price of seeing past the Horizon

Escape requires a channel that touches W: a noisy welfare query that returns, at chosen tilts up to a cap Φ_q, a measurement of welfare plus noise. The deployment-relevant target is the realized gap at deployment pressure Φ_d. Convention, used throughout: σ^2 denotes channel noise in lower bounds (converses); achievability carries $\tilde{\sigma}^2 = \bar{\sigma}^2 + B^2$, where $\bar{\sigma}^2 = \sigma^2 + \sup\_s \mathrm{Var}(W|u{=}s)$ is the channel noise plus the conditional-welfare variance, and the additive B^2 is the range bound on the conditional welfare mean $w(s)$ across the reweighted batch ($w(s) \in [-B, B]$), distinct from the conditional-variance contribution $\sup\_s \mathrm{Var}(W|u{=}s)$ already inside $\bar{\sigma}^2$.

Hindsight is cheap. When the query cap reaches deployment pressure, testing whether the gap is zero against at least δ costs a number of queries that scales as $\tilde{\sigma}^2\delta^{-2}(1 + \min(\Phi\_d \cdot \mathrm{osc}(u), K\_g)) \cdot \log(N\_grid/\alpha)$, with no χ^2 overlap penalty.

Foresight is priced. When the query cap falls short of deployment pressure, certification from a single low-pressure batch by self-normalized importance sampling succeeds on the standing bounded class at a budget governed by the overlap. No clipping is needed on that bounded class. The unbounded Gaussian-bulk idealization is kept as a labeled asymptotic regime only; its clipping constants are not claimed here until a separate unbounded truncation clause is supplied.

$$\chi^2 + 1 = Z(2\Phi\_d - \Phi\_q)\, Z(\Phi\_q) \,/\, Z(\Phi\_d)^2,$$

where Z is the proxy's moment generating function. The χ^2 overlap and the one-sided undershoot are classical (Agapiou et al. 2017; Chatterjee–Diaconis 2018); what is new here is the channel-level certification framing. A lower bound matching in functional form, via a Lipschitz-ramp two-point pair, an adaptive divergence decomposition, and Le Cam's method (Le Cam 1986; Tsybakov 2009), recovers the overlap dependence up to a $(\Phi\_d - \Phi\_q) \cdot \mathrm{osc}(u)$ band-localization residual that divides the bound and can grow with deployment strength, under a named band-dominance hypothesis with instance constant c_geo. In the worked geometries below, c_geo is about 1.36 in the atomic measured geometry and about 14.8 in a log-concave thin-edge geometry, and it can grow with edge thinness. The residual is a multiplicative factor of order $(\Phi\_d - \Phi\_q) \cdot \mathrm{osc}(u)$ that divides the lower bound, so the matching is tight only where this factor and c_geo stay bounded; absolute-constant matching is open.

The regime spectrum (now a lemma). The growth of the overlap as deployment pressure outruns the query cap is set by the thinness of the proxy's upper edge, and is now established as a numbered regime-spectrum lemma in Technical Appendix, Lemma 5.1, at a fixed query cap as deployment pressure grows, with explicit constants and error terms, via de Bruijn / Bingham–Goldie–Teugels Tauberian asymptotics. Indexing the upper edge on a single regular-variation axis:

- Atomic top (an atom of mass p^* at s^*): the overlap converges to a constant, $e^{-\Phi\_q s^*} Z(\Phi\_q)/p^*$, which is ≥ 1 and strictly > 1 once any non-atomic background is present.
- Continuous polynomial edge (edge density $\propto t^\beta$ near s^*, $\beta \geq 0$): the overlap grows like $\Phi\_d^{\beta+1}$. The box edge $\beta = 0$ grows linearly, the cheapest growing branch, and distinct from the atom. The fixed-cap constant carries the exact $Z(\Phi\_q)$; the cruder $(2\Phi\_q)^{-(\beta+1)}$ form is only its $\Phi\_q \to \infty$ reduction and is quantitatively off at the small caps actually used.
- Essential edge (density $\propto \exp(-c\, t^{-\rho})$, $\rho > 0$): the overlap is stretched-exponential of rate $\rho/(\rho+1) < 1$, not rate ρ. Within the bounded class the rate is always below 1, so the full exponential is unreachable here; the kernel's prototype $e^{-c/(s^*-s)}$ is $\rho = 1$, giving $e^{-\Theta(\sqrt{\Phi\_d})}$.
- Gaussian-bulk idealization (unbounded support, outside the standing bounded class, stated as a labeled limit): the overlap is the full Gaussian-exponential $e^{s^2\_\nu(\Phi\_d-\Phi\_q)^2}$.

One quantifier matters and is stated as such: these are fixed-query-cap asymptotics as deployment pressure grows.

Under joint scaling Φ_q = rΦ_d the polynomial overlap saturates to the pure constant (r(2−r))^{−(β+1)}. The dramatic thin-edge regimes are therefore asymptotics, not instantiated by the measured instance below, which the deep-tilt experiment of §6 finds empirically indeterminate (a non-Gaussian edge signal that does not resolve into a clean regime), not a clean affordable branch. The lemma sharpens the achievability budget (it turns the abstract overlap multiplier into explicit edge-regime budgets); it does not remove the band-localization residual of the converse, which is a property of the Le Cam two-point geometry, not of the edge asymptotics.

One-sided visibility. The overlap factor is a function of the proxy law, so in the ideal proxy-telemetry model the loop's data prices the channel it cannot replace. At finite sample, the valid certificate must account for the denominator term Z(Φ_d)^2. Lower confidence sequences on the numerator factors Z(2Φ_d−Φ_q) and Z(Φ_q), even combined with the deterministic upper handle Z(Φ_d) ≤ exp(Φ_d s*), give a lower bound capped at 1 and therefore certify nothing. The proposed realized-run certificate uses the product decomposition (Appendix C, Corollary 3Z(iii)), whose algebraic identity is exact: writing ν_Φ for the Φ-tilted proxy law ($d\nu_\Phi \propto e^{\Phi u}\, d\nu$) and Δ = Φ_d − Φ_q > 0,

$\chi^2 + 1 = E_{\nu_{\Phi_d}}[e^{\Delta u}] \cdot E_{\nu_{\Phi_d}}[e^{-\Delta u}]$,

a product of two means under the realized deployment tilt. The proposed certificate takes betting / empirical-Bernstein lower confidence sequences (Waudby-Smith and Ramdas 2024) on each factor, each at level α/2 (the α-split: α/2 per factor, so the product is intended to be anytime-valid at level α); caps $e^{\Delta u}$ at any chosen constant for the capped-mean lower bound, so no known s* is required; uses the bounded second factor (by 1 when u ≥ 0, by $e^{\Delta\|u\|_\infty}$ in general, which the standing bounded class supplies); and multiplies the two lower bounds. On the witness instance it recovers $\chi^2+1 = 1.97$, and un-hit tail mass manifests as conservative silence, a lower bound near 1, not invalidity, consistent with the degeneracy §6 reports. The available budget multiple is defined against the Section 4 self-normalized importance-sampling certification budget for the estimator family used here, of order $\tilde{\sigma}^2\delta^{-2}(1+\chi^2)\cdot\log(N_{grid}/\alpha)$: if the certified lower bound on $1+\chi^2$ already prices that budget above the queries actually affordable, the budget is certified insufficient. Scope, stated plainly: this proposed certificate applies only to realized-run pricing, where deployment-tilt samples exist by definition of the loop's own telemetry, matching the ν-identifying scope of Cor. 3Z; prospective pricing of a not-yet-realized deeper tilt from capped-tilt batches has no exhibited non-vacuous certificate and is open. This product-form certificate is stated in Appendix C, Corollary 3Z(iii), and should be treated as provisional until human mathematical referee sign-off. Diagnostics such as effective-sample-size collapse, PSIS failure, and an uninformative lower confidence sequence can also expose undercoverage in realized runs; in the sufficiency direction, a two-point pair planted on the un-sampled upper tail breaks any data-driven upper bound that omits an explicit allowance for un-hit mass. The only sufficiency certificate retained here is the a-priori worst case, using a known proxy bound. Sharper data-adaptive allowance buy-down is open. Self-certification is not merely unreliable. It is structurally one-sided, the Horizon again, now in the pricing.

The quantitative floor. Sharpening the converse, any m-query protocol capped at Φ_q has a Le Cam deficiency (against a deployment-strength experiment that separates the worlds) floored at $(1-\nu_{\Phi_q}(S_t))^m - \varepsilon$, which decreases in m only at the rate set by the top-band mass the cap can reach, which the cap keeps small exactly where deployment pressure concentrates. Empirically this appears as a one-sided undershoot, and §6 shows it survives scale: even at n = 300,000 with temperature-tilted sampling, the deepest partition-function term has effective sample size near 1 and is carried roughly 99.99% by a single completion, so the overlap estimate is formally reliable only through Φ_d = 4 and its anytime-valid lower confidence sequence is uninformative (identically zero).

5. The audit hierarchy

Oversight by a grader is the realistic middle case between a blind channel and a gold one. Model a grader by a welfare correlation ρ: its verdict carries a ρ-fraction of true welfare plus independent noise, within a Gaussian-or-degraded class. Concretely, the canonical member is $V = \rho\cdot W(a) + \sqrt{(1-\rho^2)}\cdot\sigma_0\cdot\xi$ with ξ standard Gaussian; the class is closed under welfare-independent Gaussian-conditional attenuation and post-processing. Then the foresight lower bound holds verbatim with channel noise σ^2 replaced by $\sigma_{eff}^2(\rho) = (1-\rho^2)\sigma_0^2/\rho^2$, where σ_0 is the grader's intrinsic noise scale. There are two different matched loci. If $\sigma_0 = \sigma$, the lower-bound budget carries the multiplier $(1-\rho^2)/\rho^2$ relative to the σ-noise query model. If $\sigma_{eff} = \sigma$, the grader reproduces the §4 σ-noise

query model exactly, with no extra multiplier. At $\rho = 0$ the grader is welfare-blind and the Horizon returns; as ρ grows the budget falls. The class boundary is central: a grader whose noise variance itself depends on welfare, or whose noise is merely sub-Gaussian rather than Gaussian, can break the converse, so the result is stated for the modeled class and not for "any correlated grader." Noiseless binary outcome checks (code-runs, unit tests, pass/fail verdicts) are also outside the modeled class; they may be cheaper to inform than the $(1-\rho^2)/\rho^2$ budget suggests, but only the channel-shape-free un-hit-mass floor of §4 applies to them.

This is where the Horizon becomes a hierarchy. The grader's own ρ is non-identifiable from fixed-kernel data, so an unverified grader, whatever its verdicts say, confers no certification, certifying a grader's informativeness is itself a welfare-touching measurement, one level up, and that certifier's credentials need a certifier above it in turn. Within the (OC) family and the modeled ρ-grader class, each certifier that is itself (OC) requires a welfare-touching certifier one level up; terminating the ladder requires stepping outside that closure, a welfare-touching channel, including a noiseless outcome check where the deficit is outcome-defined and such a check exists. Scalable-oversight proposals such as debate (Brown-Cohen et al. 2024) inherit the same condition: to ground the ladder they must, somewhere, touch welfare rather than only the trajectory, which is exactly the priced escape of §4.

6. A worked instance

On a reward-model substrate (GPT-2-medium completions scored by two DeBERTa reward models, one as proxy and one stipulated as gold), the trajectory is rising at every reachable pressure, the benign branch, measured rather than assumed (within-sample tilting at $n = 128$ is bounded by best-of-128 selection, so a population peak beyond the reachable band is not excluded, the blindness theorem's point in vivo, not a design deficiency). The certification price is computed from proxy telemetry alone; the definitive measurement, detailed below, scales the released sample to $n = 300{,}000$ with temperature-tilted sampling (50 prompts; 4,000 base completions per prompt at $T = 1.0$, plus 1,000 each at $T = 1.5$ and $T = 2.0$; a duplicate concurrent sampler was deduplicated by the (prompt, temperature, sample) key to exactly 300,000 verified rows; model revisions pinned in the repository) and prices the overlap across deployment pressures $\Phi\_d \in [3, 15]$ at query cap $\Phi\_q = 2$. The one-sidedness is visible directly: at query strength the partition-function estimate is stable across subsampling, while at deployment strength it drifts by orders of magnitude as the rare upper tail is hit, with every lower confidence band intact. A second, separate illustration fits the Agrawal et al. (2025) LLM double-auction collusion experiment (five frozen-LLM sellers and five buyers) as a tilt path; this is a per-arm fit, not a structural validation, and its misses are reported.

The deep-tilt experiment: the missing mass survives scale. The released $n = 6{,}400$ sample was too sparse to identify the regime-spectrum edge exponent: the deepest moment-generating argument was carried by a single completion. The larger deep-tilt-importance-sampled draw confirms the diagnosis rather than dissolving it. Scaled to $n = 300{,}000$ with temperature-tilted sampling, the overlap estimate is nearly flat: $1 + \hat{\chi}^2 = 1.024$ at $\Phi\_d = 3$, saturating near 1.027 for $\Phi\_d \geq 5$. Read naively, a flat overlap would place the instance in a cheap, atom-like branch. It cannot be read that way. The PSIS-k reliability criterion is met only through $\Phi\_d = 4$; beyond it the importance estimate is formally unreliable. And even where it is formally reliable, the estimate is degenerate: the effective sample size of the deepest partition-function term is near 1 throughout, a single completion carries roughly 0.9999 to 1.0000 of its mass, and the anytime-valid lower confidence sequence is identically zero. The flat overlap is therefore an extreme-order-statistic object, not a population estimate, and its apparent cheapness is exactly the artifact the theorem predicts: one dominating completion makes the partition function look atom-carried, while the un-hit upper-tail mass that could inflate the true overlap without bound is never sampled. This is Corollary 3CC and the qualitative one-sidedness of 3Z in vivo, and it survives a forty-seven-fold increase in sample size and the addition of tilted sampling. (Tilted sampling did move the estimate, from the base-only 1.58-to-4.20 dense read down to the flat 1.027 curve; but the single-completion domination persists in both, so neither is the population overlap, and the gap between the two is itself the instability the missing-mass warning names.)

The edge classification sharpens but stays indeterminate. On the base sample the finite-endpoint density estimator now returns $\hat{\beta} \approx 10.22$ and separates from a Gaussian null (null median 5.55, 95% interval [4.48, 7.11]), a genuine non-Gaussian upper-edge signal where the $n = 6{,}400$ read had none. But the peaks-over-threshold (GPD) shape parameter ξ is sign-unstable across thresholds: $\xi = -0.060$ at the top 10% (bounded-edge sign), turning posi-

tive at +0.029, +0.056, and +0.063 at the top 5%, 2%, and 1% (heavy-tail sign), in conflict with the bounded support the construction assumes. The reliable-region overlap fit, separately, returns a leading exponent $\beta \approx -0.99$, the atom endpoint. These do not cohere into a clean classification. The honest verdict, which the report records, is a non-Gaussian finite-endpoint signal with unstable POT/GPD behavior: neither a clean bounded-polynomial nor a clean essential edge, but empirically indeterminate. Placing the instance on the regime-spectrum lemma's β-axis is not yet warranted; the edge exponent remains behind the Horizon.

The diagnostics work when there is something to see. Synthetic positive controls confirm the pipeline detects thin edges when they are present, so the real-proxy indeterminacy is a property of the data, not a failure of the method. A known polynomial edge ($\beta = 2$) is recovered (known-endpoint $\beta \approx 1.95$, finite-endpoint $\beta \approx 1.89$, separating from the Gaussian null). A known essential edge is detected, its known-endpoint exponent rising with depth from ≈ 3.86 at the top 10% to ≈ 5.46 at the top 1%, the signature of an effectively increasing edge, and it separates from the null. A Gaussian control returns finite-endpoint $\beta \approx 4.33$ and does not separate. The pipeline discriminates the cases it is built to discriminate; the real proxy does not present a clean one.

What this settles. Scaling the reward-model experiment to n = 300,000 and adding tilted sampling does not remove the one-sided missing-mass pathology. The reliable region stops at Φ_d = 4, and even there the deepest term is a single completion, so the high-pressure overlap is an extreme-order-statistic object, not a stable population curve. The edge is a non-Gaussian signal but POT-unstable, so it is reported as indeterminate, not as affordable or Gaussian-indistinguishable; those were prior expectations, not measured results. Internal proxy telemetry can show that the certification problem is undercovered and one-sided. It cannot, by itself, certify high-pressure sufficiency without explicit tail assumptions or a genuine welfare-touching channel. The Horizon holds at scale, on the paper's own largest measurement.

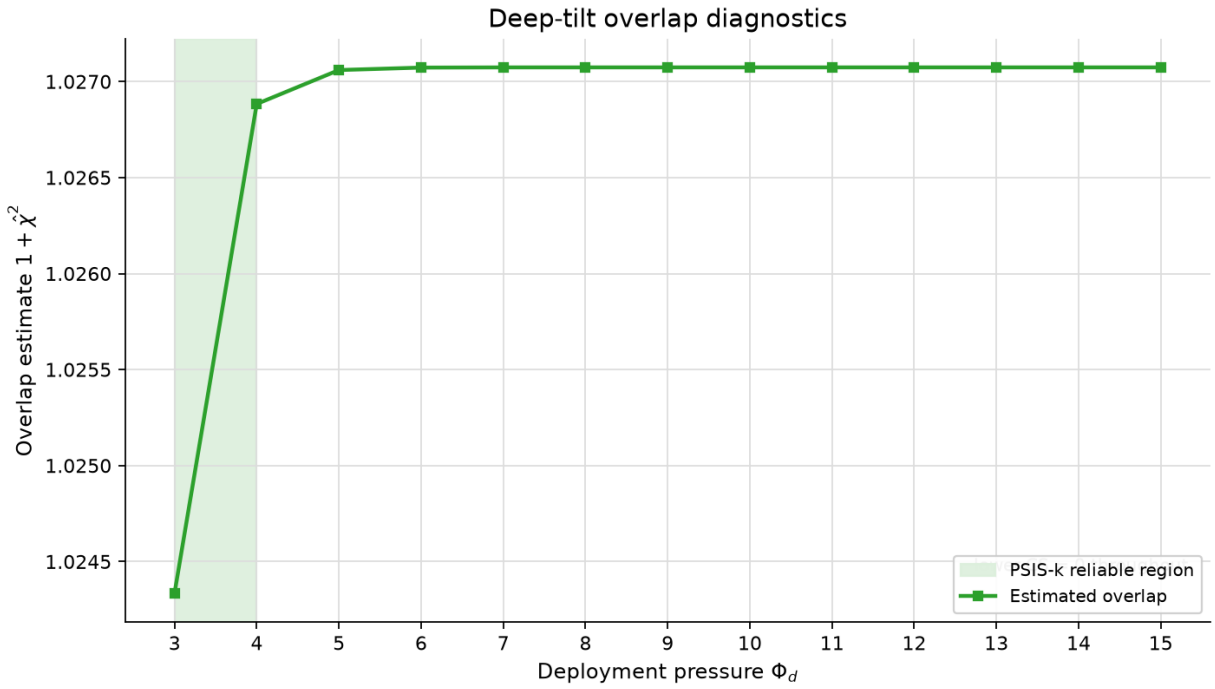

Figure 1: The deep-tilt overlap experiment (n = 300,000, temperature-tilted sampling). The overlap $1 + \hat{\chi}^2$ versus deployment pressure Φ_d at query cap Φ_q = 2 is nearly flat (≈ 1.024 rising to ≈ 1.027), but the diagnostics make that flatness uninformative: the effective sample size of the deepest partition-function term is ≈ 1 and a single completion carries ≈ 1.0 of its mass throughout, and PSIS-k reliability holds only through Φ_d = 4. The flat curve is an extreme-order-statistic object, not a population overlap; its apparent affordability cannot be certified. This is the one-sided missing-mass pathology of §4, at scale.

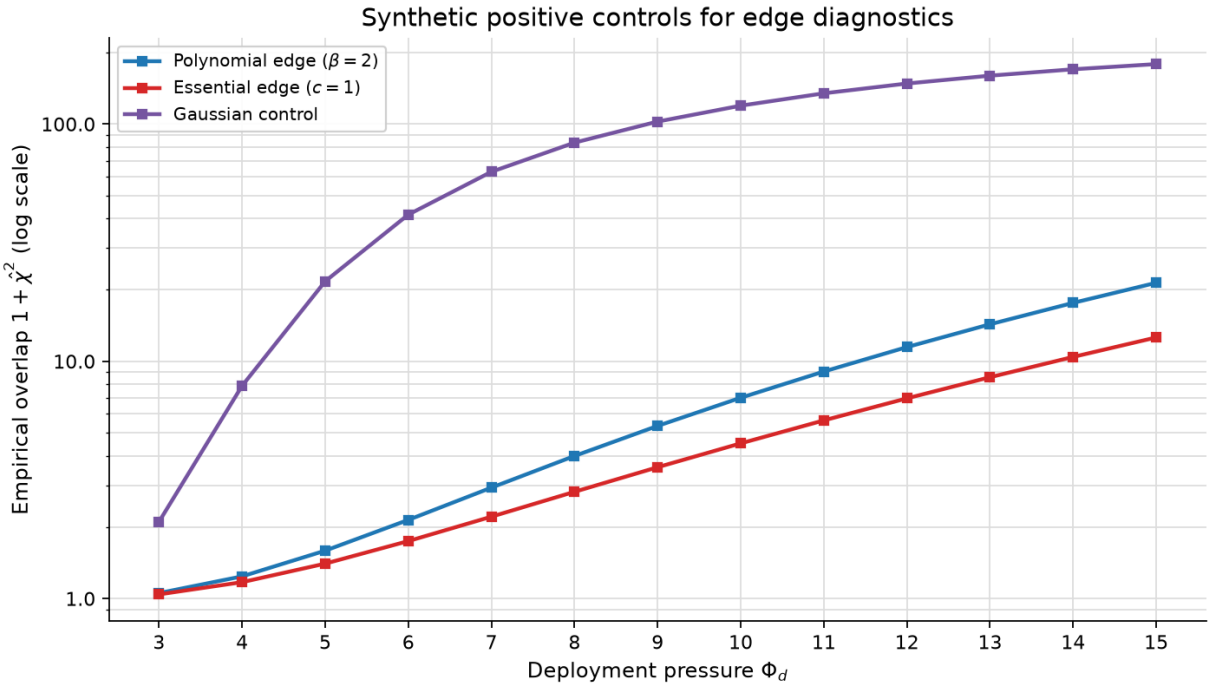

Figure 2: Synthetic positive controls for the edge diagnostic. The same pipeline recovers a known polynomial edge ($\beta \approx 2$) and detects a known essential edge (known-endpoint exponent rising 3.86 to 5.46 with depth), both separating from a Gaussian null, while a Gaussian control does not separate. The pipeline sees thin edges when they are present, so the real proxy's indeterminate verdict reflects the data, not a method failure.

7. The meta-horizon: pricing the governance channel

A separate mechanism-design companion in preparation explores when, under stated substrate-dependent conditions, no mechanism can satisfy incentive compatibility, individual rationality, budget balance, delivering share α of attainable surplus (accounted net of operator information rents), and safety certification at the pressure cap. The result is summarized here, not claimed as a theorem of this paper.

Under incentive compatibility, participation, and budget balance, requiring delivery of a target share of attainable gains (accounted net of operator information rents) forces any mechanism to license activity above a floor; the binding axiom is the deliverable-gains accounting, not the classical IC/IR/BB triple, which alone admits the efficient allocation in this environment. The companion uses the Myerson rent identity together with an explicit residual claimant (in the budget-breaker tradition of Holmström 1982) (plus firewall axioms separating the mechanism's information set from welfare, an agent-side firewall that excludes Crémer–McLean correlated-types escapes by assumption, a real concession in human settings, and a mechanism-side firewall that holds whenever surplus accounting is proxy-denominated rather than welfare-denominated). Worst-case alignment caps admissible pressure at a one-sided certified ceiling. When the activity floor exceeds the ceiling, the companion shows no mechanism satisfies all five desiderata (incentive compatibility, individual rationality, budget balance, delivering share α of attainable surplus, and safety certification at the pressure cap). This is conditional and substrate-dependent, a joint condition on the proxy law and the activity-pressure coupling (plus the firewall axioms above; see companion for the full hypothesis list), measured nonempty on one substrate and empty on another at the same coupling. It is a condition to check, not a law of nature.

A second impossibility sits at the exclusion-boundary corner, price- and edge-regime-independent though not unconditional: within the companion's symmetric regular interior-exclusion environment, at the full-share corner $\alpha = 1$ the threshold rule reaches the Myerson exclusion type, the marginal virtual surplus at the floor is zero, and internal financing of any audit fails by rents alone, at any audit price, in any edge regime. Outside that environment the corner result is not claimed.

This is the Horizon one level up. An institution that tries to audit itself is a system reading its own telemetry: if that self-audit channel satisfies (OC) (its outputs a fixed function of internal telemetry with no welfare-dependent path, an assumption far less automatic for institutions, with their whistleblowers and leaks, than for a frozen model) then Lemma 1's logic recurs, and its self-audit cannot certify what only a channel from outside can. The escape is exogenous outcome-contact, a welfare-touching channel the institution cannot compute from its own trajectory; external audit is one realization, and a purely document-reading external auditor, itself (OC), is not. That channel cannot be self-certified from inside the institution, and beyond a regime-dependent solvency frontier the achievability-priced version cannot be internally financed. In the impossibility region, any institution preserving non-trivial gains must import exogenous outcome-contact, by whatever instrument; the one inadmissible shape is self-certified, trajectory-only audit.

The practical reading is direct. Production AI systems optimize proxies, and the field's instinct, when something looks wrong, is to add monitoring. This result says that instinct is structurally limited: more of the telemetry a system already has cannot see its own crossing, because that telemetry is a function of the trajectory and the welfare deficit never entered it. Only a channel that touches welfare can, and that channel is expensive in exactly the open-ended regimes where frontier optimization lives. The welfare-touching channels (gold evaluations, outcome audits, red-teams in unfamiliar regimes, deployment-time interventions) are where the field's safety toolkit already lives; what is structurally limited is reading the crossing from the system's own training-time data.

8. Status and open problems

What is verified. The covariance identity, the overlap functional, the Pinsker constants, the Gaussian channel divergence, and the graded-oversight multiplier reproduce under symbolic and high-precision computation (sympy/mpmath). The bounded-class foresight achievability is the claim retained here; the unbounded clipping constants remain outside the certified constant stack and are left for later work. The regime-spectrum constants are established in Technical Appendix, Lemma 5.1, subject to human mathematical refereeing. The constructions and proof sketches have been cross-checked under independent verification across the relevant disciplines; the remaining proof package still needs human mathematical review.

What is open. The proofs have not been refereed by a human mathematician; that remains the binding gate, and the realized-tilt product-form certificate of §4 is inside it. The finite-sample insufficiency certificate is proposed for realized-run pricing via the product form and remains provisional pending review; a non-vacuous prospective certificate (pricing a not-yet-realized deeper tilt from capped-tilt batches) remains open, as does sharper data-adaptive allowance buy-down in the sufficiency direction. An absolute-constant version of the matching converse (overlap necessity) is open; the regime-spectrum lemma is orthogonal to it. And the mechanism-design crossing of §7 is established for a stated, substrate-dependent class.

Data and code availability. The public replication repository is https://github.com/darfaz/goodhart-gap-experiment, with an archived Zenodo replication bundle at DOI 10.5281/zenodo.20722572 (concept DOI; current release v0.1.0 archived at 10.5281/zenodo.20722573, record 20722573, checksum `md5:d5db5dd3afe76e83ad1562a196929bbf`). Full Goodhart proof details are included in the technical appendices below. The deep-tilt extension reported here uses the same experimental code path and is represented in this paper by the embedded Figure 1 and Figure 2 assets; a public replication update containing its report, tables, and source data is pending. Collaboration on the open mathematical items is welcome.

Technical Appendices

The following appendices fold in the formal proof kernel so the paper is self-contained. They retain the theorem labels used in the main text; remaining caveats are mathematical scope conditions.

3. The gap: factorization and trichotomy

Lemma 1 (master identity). With μ_0 a probability measure and u, W bounded, Z is entire, T_w is real-analytic on $\mathbb{R}$, and T_w′(Φ) = Cov_{π_Φ}(u, W). Welfare rises under pressure exactly while proxy and welfare remain positively correlated under the current tilted measure; optimization changes the measure, and the measure change is what kills the correlation — Goodhart's law as an identity.

Lemma 2 (factorization). T_w(Φ) = $\int$ w(s) e^{Φs} ν(ds) / Z(Φ). The trajectory depends on W only through (ν, w).

Theorem 1 (trichotomy and the Goodhart gap). Assume (G3): L := lim_{Φ→∞} T_w(Φ) exists (sufficient: w has a limit at s*; or ν has an atom at s*, in which case no further regularity is needed). Let Δ := sup_{Φ≥0} T_w(Φ) − L ≥ 0; when the supremum is attained, Φ* denotes the smallest maximizer. Then: (a) Δ = 0 — aligned at the top; pressure never costs welfare relative to its limit. (b) if Cov_{μ_0}(u, W) > 0 (G2) and Δ > 0 — the supremum is attained at finite interior Φ* with Cov_{π_{Φ*}}(u, W) = 0; the trajectory never exceeds the peak thereafter and gives back Δ, by definition the peak-to-limit decline. No unimodality or monotone decline past Φ* is claimed (under strict TP_2 those follow classically; we do not need them). (c) if Cov_{μ_0}(u, W) < 0 — welfare falls below baseline from the first unit of pressure. These cases exhaust the behavior except for two boundaries: Cov_{μ_0}(u,W) = 0 (degenerate; the interior-supremum argument still applies when Δ > 0, without the Φ* > 0 conclusion), and the overlap of (a) with (c) when an initial decline approaches the limit from below. Scope: Δ > 0 is substantive, not generic — for W = u + εv with v ∈ C^2 near an interior nondegenerate maximum of u that is the unique global maximizer (μ_0-essentially, with Laplace localization: sup of u outside every neighborhood strictly below u*; finite-dimensional chart) and μ_0 with locally C^1 (or Lipschitz) positive density at the maximizer, Δ = 0 for ε small; proxies that track welfare to the top are safe and the theorem says so. Boundary maxima, densities thin or vanishing at the maximizer, and proxies with multiple global maximizers (ties at the top) are outside the claim — a delimitation consistent with the paper's edge-thinness theme. (G3) is necessary: without limit regularity the trajectory can oscillate forever. (Proofs: Appendix A.)

4. The impossibility

Theorem 2 (non-identifiability). Fix (μ_0, u) and any observation process satisfying (OC). The observation law is a functional of the trajectory law generated by (μ_0, u) and the fixed channel kernel, and is invariant to W (in proxy-score-only reductions this law factors through ν). Consequently all welfare functions over a fixed (μ_0, u) generate identical observation laws; pairs agreeing in conditional mean away from the proxy optimum but differing in its limit at the optimum realize different Δ, hence different pairs (Δ, Φ*) whenever pressure actually

moves mass toward the top. In particular, on nondegenerate proxy laws with arbitrarily small baseline mass in top bands, worlds with $\Delta = 0$ and Δ growing with the welfare bound B (always $\Delta \le 2B$) are observationally identical; degenerate laws where the baseline is already concentrated at the proxy optimum are excluded from this large-gap clause. Non-identification of the peak location between distinct finite interior values is not exhibited by this construction. No test, estimator, eval, or stopping rule computed from (OC) data distinguishes the exhibited worlds above chance, in any sample size, forever.

Remarks. (1) No adversary, no computational assumptions, no noise floor: this is the geometry of what the channel can express — distinct from cryptographic undetectability of planted behavior (Goldwasser et al. 2022). (2) Scope of the interpretability implication: the two worlds differ only in W, a latent outcome function with no causal path into the system, so weights and activations are identical and uninformative in principle — but only in the regime where the welfare deficit is outcome-defined (code that reads well versus code that runs; engagement versus downstream harm). When misalignment is encoded in the mechanism itself (deception, sandbagging), the construction does not apply and Theorem 2 is silent; interpretability targets exactly that complementary regime. (3) The result is relative to data generated under the deployed regime; channels that intervene violate (OC) and are the escape — consistent with the econometric finding that Lucas-critique violations are detectable only through regime variation.

5. The price of the W-channel

Escape channels exist; this section prices them. The monitor gains a noisy welfare channel: queries at chosen tilts $\Phi_i \le \Phi_q$ return (s_i, Y_i), $Y_i = W(a_i) + \sigma\xi_i$. Deployment-relevant target: the realized gap $G_w(\Phi_d) = \max_{\{\Phi \le \Phi_d\}} T_w(\Phi) - T_w(\Phi_d)$. Lower bounds transfer to Theorem 1's Δ immediately (certifying $\Delta \le \delta$ is harder); upper bounds transfer once Φ_d exceeds the (G3) convergence scale. Statements 3R-3Z give the channel-pricing package; Proposition 3G and the realized-tilt product certificate remain subject to human mathematical review. The χ^2 overlap functional and the one-sided undershoot have classical ancestry — the e^{KL} importance-sampling sample-size law of Chatterjee and Diaconis (2018) and the χ^2/second-moment characterization of Agapiou, Papaspiliopoulos, Sanz-Alonso, and Stuart (2017); what is new here is the channel-level certification framing (§7). Convention throughout: σ^2 denotes channel noise in lower bounds (converses); upper bounds (achievability) carry $\tilde{\sigma}^2 := \bar{\sigma}^2 + B^2$, where $\bar{\sigma}^2 = \sigma^2 + \sup_s \mathrm{Var}(W|u{=}s)$ — the additive B^2 is the conditional-mean fluctuation that importance reweighting exposes (the estimand w(s) itself ranges over $[-B, B]$ across the reweighted batch), and it is not optional.

Theorem 3R (hindsight is affordable). For $\Phi_d \le \Phi_q$: testing $G = 0$ vs $G \ge \delta$ over all bounded measurable w costs $n \le C\tilde{\sigma}^2\delta^{-2}(1 + \min(\Phi_d \cdot \mathrm{osc}(u), K_g)) \cdot \log(N_{grid}/\alpha)$, via batch-window importance reweighting on a grid of size $N_{grid} = O(\Phi_d \|u\|_\infty B/\delta)$, where K_g is the greedy e^{1/2}-cover count of $[0, \Phi_d]$ (Appendix C). Proven unconditionally $K_g \le \lceil \Phi_d \cdot \mathrm{osc}(u) \rceil + 1$; empirically $K_g \approx \int_0^{\Phi_d} \sqrt{(2 \cdot \mathrm{Var}\{\nu_\Phi\}(s))} d\Phi \le \sqrt{(2\Phi_d \cdot \mathrm{osc}(u))}$ (order-equivalence numerically verified, not proven); the linear factor $\Phi_d \cdot \mathrm{osc}(u)$ is the Popoviciu worst case, not a real information cost.

Theorem 3P (foresight lower bound). For $\Phi_q < \Phi_d$, over the class of w bounded by B and L-Lipschitz near s^*, with the feasible band set $\mathcal{T} = \{t \le t_0 : \nu(S_t) \le \tfrac{1}{4}\nu_{\{\Phi_d\}}(S_{t/2})\}$; ($t_0 > 0$ an arbitrary fixed band cap) $Lt \ge 4\delta/\nu_{\{\Phi_d\}}(S_{t/2})\}$; $\nu_{\{\Phi_d\}}(S_{t/2}\} \ge 4\delta/B\}$ ($\mathcal{T} \ne \varnothing$ is an instance hypothesis, not automatic; checkable sufficient condition: an atomic or sufficiently concentrated upper edge with top-band mass $p_d := \lim_{t\to 0} \nu_{\{\Phi_d\}}(S_{t/2}\}) > 0$, baseline concentration $\nu(S_t) \le p_d/4$ on the relevant bands, and $\delta \le (p_d/4) \cdot \min(Lt_0, B)$, under which $\mathcal{T} \supseteq [4\delta/(Lp_d), t_0]$; thin no-atom edges where $\nu_{\{\Phi_d\}}(S_{t/2}\} \to 0$ as $t \to 0$ can leave $\mathcal{T} = \varnothing$ for every δ and need a separate instance check — Appendix C): any test with both errors $\le 1/3$ requires $n \ge (1/36)\, \sigma^2 \delta^{-2} \cdot \sup_{t\in\mathcal{T}} \nu_{\{\Phi_d\}}(S_{t/2}\}^2/\nu_{\{\Phi_q\}}(S_t)$. Proof by a Lipschitz-ramp two-point pair sharing all (OC) data, adaptive divergence decomposition, Le Cam; the bound concerns the gap functional, not the branch — a common rising component added to both worlds places the same pair in branch (b) with the KL unchanged (Appendix C).

Theorem 3P′ (foresight upper bound; bounded class). On the standing bounded class, self-normalized importance sampling from a single $\nu_{\{\Phi_q\}}$ batch certifies $G \le \delta$ vs $\ge 2\delta$ at $n \le C\tilde{\sigma}^2\delta^{-2}[(1 + \min(\Phi_q \cdot \mathrm{osc}(u), K_g(\Phi_q))) + (1 + \chi^2(\nu_{\{\Phi_d\}} \| \nu_{\{\Phi_q\}}))] \log(N_{grid}/\alpha)$ — stage (i) is Theorem 3R on $[0, \Phi_q]$, inheriting its factor — with the closed form $\chi^2 + 1 = Z(2\Phi_d - \Phi_q)Z(\Phi_q)/Z(\Phi_d)^2$. No clipping is needed on this bounded class. The unbounded Gaussian-bulk idealization in the regime spectrum is not covered by this theorem's printed constants; a valid

unbounded truncation clause requires a separate clipping budget and is left for later work. Lower and upper bounds match up to a multiplicative O((Φ_d−Φ_q)·osc(u)) (attained) plus a band-localization residual in the current lower/upper comparison: clipped/truncated estimators narrow it in computed geometries but do not close it there, and the multi-band refinement of the lower bound recovers only the feasible-band sum (a bounded factor at the geometries computed); what remains open is whether a different proof technique collapses it.

The regime spectrum. Within the standing bounded class, the price is decided by the thinness of ν's upper edge. For a polynomial edge with density asymptotic to c_ν(s*−s)^β near s*, let a = β + 1 and C = c_νΓ(a) (c_ν = c_0 of Definition 5.1(3) under the second-order hypothesis). At fixed query cap Φ_q, Watson's lemma gives the corrected public display $\chi^2+1 \sim 2^{\{-a\}e}\{-\Phi_q s^*\}Z(\Phi_q)C^{\{-1\}\Phi_d}a$ as Φ_d → ∞ — polynomial in deployment strength, with the exact fixed-cap constant carrying Z(Φ_q). The cruder $(2\Phi_q)^{\{-a\}\Phi_d}a$ form is only the further large-Φ_q reduction and is not used for small-cap constants. An essential edge (density ~e^{−c/(s*−s)}) gives the stretched-exponential e^{Θ($\sqrt{\Phi_d}$)}; the Gaussian-bulk price e^{$s^2_\nu(\Phi_d-\Phi_q)^2$} is the unbounded idealization (outside the standing assumptions — the regime of effectively unbounded learned scores, stated as a labeled limit, not an instance of the bounded theory). Evaluate-at-strength is information-theoretically forced exactly for open-ended objectives whose optimum the tilt path has not saturated — the reward-model overoptimization habitat — and merely prudent where deployment saturates an attained maximum. Quantifier and scope: the displayed closed forms are fixed-Φ_q (query cap held) asymptotics as Φ_d → ∞; under joint scaling $\Phi_q = a_0\Phi_d$ the polynomial form saturates to the constant $[(2-a_0)a_0]^{-(1+\beta)}$ (= Lemma 5.1(E)'s $(r(2-r))^{-(\beta+1)}$ with $r = a_0$), so "polynomial/superpolynomial in deployment strength" is the held-query-cap statement. The thin-edge superpolynomial branch is an edge-asymptotic, not instantiated as such: the §6.3 measured instance's edge is, on the definitive deep-tilt run (n = 300,000, tilted sampling), a non-Gaussian finite-endpoint signal with sign-unstable POT/GPD behavior — empirically indeterminate (§6.3(iii)), neither a clean affordable/atomic branch nor a clean thin edge; the earlier n = 6,400 read was insufficient, and the deep-tilt overlap is flat (~1.027) but single-completion-dominated, so the population price is not claimed. These asymptotics are stated formally in Lemma 5.1 below, with explicit hypotheses and error terms.

5.1 The regime-spectrum lemma (formal statement and proof outline)

Technical status. All limits are deterministic limits of Borel integrals; the essential supremum s* is the usual ν-essential supremum. The closed forms and remainder terms have been reproduced at high precision on explicit test laws, but human mathematical refereeing remains an open item and gates external mathematical authority.

Notation. Let $(\mathcal{A}, \mathcal{F}, \mu_0)$ be a probability space, u: $\mathcal{A} \to \mathbb{R}$ the proxy, $\nu = u_*\mu_0$ its pushforward law, and $s^* = \text{ess-sup}_\nu\, s \le \infty$. Write $Z(\Phi) = \int e^{\{\Phi s\}}\, \nu(ds)$ and, for finite Z(Φ), $\nu_\Phi(ds) = e^{\{\Phi s\}}\, \nu(ds) / Z(\Phi)$. For $\Phi_q < \Phi_d$ the foresight overlap appearing in Theorem 3P′ is

$$1 + \chi^2(\nu_{\Phi_d} \,\|\, \nu_{\Phi_q}) = \int \left(d\nu_{\Phi_d}/d\nu_{\Phi_q}\right)^2 d\nu_{\Phi_q} = Z(2\Phi_d - \Phi_q)\, Z(\Phi_q) / Z(\Phi_d)^2.$$

Quantifier convention. Unless joint scaling is explicitly stated, all asymptotics are at fixed $\Phi_q \ge 0$ as Φ_d → ∞. In the overlap, Z(2Φ_d − Φ_q) and Z(Φ_d) have arguments tending to ∞ and are replaced by their edge asymptotics, while Z(Φ_q) is a fixed finite constant and is never replaced by an asymptotic form. The displayed O(·) remainders are non-uniform as Φ_q → ∞; that proportional regime Φ_q = r Φ_d is the separate joint-scaling corollary, part (E).

Definition 5.1 (Upper-edge classes). In the bounded case $s^* < \infty$, put $T = s^* - S$ and write the upper-edge density, when it exists, as $f^*(t) = d\nu(s^* - t)/dt$ for t ↓ 0. Index the classes on a single regular-variation axis:

1. Atomic top (β = −1 endpoint). $\nu(\{s^*\}) = p^* > 0$.
2. Regularly varying continuous edge ($\beta \in (-1, \infty)$). For some β > −1, c > 0, and slowly varying L, $f^*(t) = c\, t^\beta\, L(1/t)(1 + o(1))$. The kernel's continuous polynomial branch corresponds to β ≥ 0. The tail mass then has index β + 1. Thus an atom is not the β = 0 case under the density-index convention; the continuous β = 0 case is a positive finite endpoint density.
3. Second-order polynomial edge. For explicit constants and a true O(Φ_d^{−1}) correction we use the stronger local hypothesis $f^*(t) = c_0\, t^\beta\, (1 + \eta\, t + O(t^2))$, β ≥ 0, as t ↓ 0. More precisely, there are t_0, M

$< \infty$ such that the remainder has absolute value at most M t^{β+2} on (0, t_0). Here $\eta \in \mathbb{R}$ is the linear edge-germ coefficient (it may be negative).

4. Essential edge, pure-power saddle form ($\beta = \infty$). For C_0, c > 0, ρ > 0, b ∈ ℝ, and η_E > 0, f*(t) = C_0 t^b exp(−c t^{−ρ})(1 + O(t^{η_E})). The index ρ > 0 is any positive number; in particular the kernel's own prototype f*(t) ≍ e^{−c/t} is ρ = 1 and is included (it is not restricted to ρ ∈ (0, 1)). This is the explicit-constant version of the de Bruijn / Bingham–Goldie–Teugels essential-edge class. In the general form exp[−ψ(1/t)] with ψ regularly varying of index ρ, the stretched-exponential rate is ρ/(ρ+1), not ρ.

5. Gaussian-bulk idealization. This is outside the standing bounded-support assumptions: S ~ N(μ, s_ν²), so Z(Φ) = exp(μΦ + s_ν² Φ² / 2).

Lemma 5.1 (Regime spectrum for the Goodhart-Gap overlap). Let Φ_q ≥ 0 be fixed unless joint scaling is explicitly stated.

(A) Atomic top ($\beta = -1$). If s* < ∞ and p* = ν({s*}) > 0, then

$1 + \chi^2(\nu_{\Phi_d} \| \nu_{\Phi_q}) \to e^{-\Phi_q s^*} Z(\Phi_q) / p^* \ (\Phi_d \to \infty)$.

Since Φ_q ≥ 0 gives $Z(\Phi_q) = \int e^{\Phi_q s} d\nu \geq p^* e^{\Phi_q s^*}$, this finite constant is ≥ 1, with equality iff p* = 1 (a pure atom); for an atom-plus-background (p* < 1) it is strictly > 1. If, in addition, the non-atomic part has a second-order polynomial edge of exponent β ≥ 0, the convergence has remainder $1 + O(\Phi_d^{-(\beta+1)})$.

(B) Continuous polynomial edge (β ≥ 0). Assume s* < ∞ and the second-order polynomial edge hypothesis. Put α = β + 1 and b_1 = α η. Then, for every fixed Φ_q ≥ 0,

$1 + \chi^2(\nu_{\Phi_d} \| \nu_{\Phi_q}) = c_\beta(\Phi_q)\, \Phi_d^\alpha \cdot (1 + (\alpha\, \Phi_q - 3\, b_1) / (2\, \Phi_d) + O(\Phi_d^{-2}))$,

where

$c_\beta(\Phi_q) = 2^{-\alpha} \cdot e^{-\Phi_q s^*} Z(\Phi_q) / (c_0\, \Gamma(\alpha))$.

The $O(\Phi_d^{-2})$ constant may depend on the fixed Φ_q and on the edge-germ bounds, and is uniform for Φ_q in any compact interval. The case β = 0 is the continuous endpoint-density case (α = 1, linear growth $1 + \chi^2 \asymp \Phi_d$); it is the cheapest growing branch, distinct from the atom β = −1 (constant). Under the weaker Karamata hypothesis f*(t) = c t^β L(1/t)(1 + o(1)), the correct fixed-Φ_q statement is

$1 + \chi^2(\nu_{\Phi_d} \| \nu_{\Phi_q}) = [e^{-\Phi_q s^*} Z(\Phi_q) / (c\, \Gamma(\alpha))] \cdot [\Phi_d^{2\alpha} / (2\Phi_d - \Phi_q)^\alpha] \cdot [L(2\Phi_d - \Phi_q) / L(\Phi_d)^2]\ (1 + o(1))$.

The closed constant above is not available for an arbitrary slowly varying factor; the power α = 1 + β survives, but multiplied by the slowly varying L(2Φ_d − Φ_q) / L(Φ_d)². If L(x) → ℓ ∈ (0, ∞) and the second-order expansion holds with c_0 = c ℓ, the displayed c_β(Φ_q) follows.

(C) Essential edge ($\beta = \infty$). Assume the pure-power saddle form f*(t) = C_0 t^b exp(−c t^{−ρ})(1 + O(t^{η_E})) with ρ > 0. Define

$p = \rho/(\rho+1) \in (0, 1)$, $K = (\rho+1)\, c^{1/(\rho+1)}\, \rho^{-\rho/(\rho+1)}$, $m = (2b + \rho + 2) / (2(\rho+1))$, $A_E = C_0 \sqrt{(2\pi/(\rho+1))} \cdot (c\, \rho)^{(2b+1)/(2(\rho+1))}$, $\varkappa = K\,(2 - 2^p)$, $C_E(\Phi_q) = 2^{-m} \cdot e^{-\Phi_q s^*} Z(\Phi_q) / A_E$, $\gamma = \min\{p, 1 - p, \eta_E / (\rho+1)\} > 0$.

Then, for every ρ > 0 and every fixed Φ_q ≥ 0,

$1 + \chi^2(\nu_{\Phi_d} \| \nu_{\Phi_q}) = C_E(\Phi_q)\, \Phi_d^m \exp\{\varkappa\, \Phi_d^p\}\, (1 + O(\Phi_d^{-\gamma}))$,

equivalently $\log(1 + \chi^2) = \varkappa\, \Phi_d^p + m \log \Phi_d + O(1)$. For the kernel's example f*(t) ≍ exp(−c/t), one has ρ = 1, p = 1/2, and $\varkappa = 2\sqrt{c}\,(2 - \sqrt{2})$, recovering $e^{\Theta(\sqrt{\Phi_d})}$. Within the bounded class p < 1 always, so the full exponential is unreachable here (it needs unbounded support, part (D)).

For a general essential edge f*(t) ≍ exp[−ψ(1/t)] with ψ regularly varying of index ρ, the de Bruijn transform gives $\log Z(\lambda) = \lambda s^* - H(\lambda) + O(\log \lambda)$ with $H(\lambda) = \inf_{t>0}\{\lambda\, t + \psi(1/t)\}$ regularly varying of index p = ρ/(ρ+1) under the usual smooth variation hypotheses; hence $\log(1 + \chi^2) = (2 - 2^p)\, H(\Phi_d)(1 + o(1))$. The statement with rate "a" is correct only if a denotes the stretched-exponential rate p (equivalently, ψ has index a/(1−a)).

(D) Gaussian-bulk idealization. If $S \sim N(\mu, s_{\nu^2})$, then, exactly and for all $\Phi_q < \Phi_d$,

$1 + \chi^2(\nu_{\Phi_d} \| \nu_{\Phi_q}) = \exp\{s_{\nu^2} (\Phi_d - \Phi_q)^2\}$.

This is a labeled unbounded-support limit, not a bounded-edge instance.

(E) Joint polynomial scaling. In the continuous regularly varying polynomial case ($\beta \geq 0$, $L \to \ell$), if instead Φ_q, $\Phi_d \to \infty$ with $\Phi_q / \Phi_d \to r \in (0, 1)$, then

$1 + \chi^2(\nu_{\Phi_d} \| \nu_{\Phi_q}) \to (r(2 - r))^{-(\beta+1)}$.

This is a pure constant: under joint scaling $Z(\Phi_q)$ is itself a large argument and is absorbed, so there is no leftover $Z(\Phi_q)$, $e^{-\Phi_q s^*}$, or $1/(c_0\, \Gamma(\alpha))$ factor. As $r \to 1^-$ the limit is 1 (equal pressures, overlap $\to 0$), and as $r \to 0^+$ it diverges, matching the fixed-Φ_q growth $\Phi_d^{1+\beta}$ of (B). In the atomic-top case the same joint limit is 1 (the p^* factors cancel).

Proof sketch. The atomic branch splits $Z(\lambda) = e^{\lambda s^*} (p^* + \int_{(0,\infty)} e^{-\lambda t}\, \nu^*(dt))$ where ν^* is the image of the non-atomic part under $t = s^* - s$; the second term is $o(1)$ by dominated convergence on the large arguments, giving $1 + \chi^2 \to p^* e^{(2\Phi_d - \Phi_q) s^*} Z(\Phi_q) / (p^{*2} e^{2\Phi_d s^*}) = e^{-\Phi_q s^*} Z(\Phi_q) / p^*$; the ≥ 1 clause uses $Z(\Phi_q) \geq p^* e^{\Phi_q s^*}$. Under a second-order polynomial edge on the non-atomic part, Watson's lemma gives the sharper $O(\Phi_d^{-(\beta+1)})$ remainder. The continuous polynomial branch localizes at the endpoint, applies Watson at a boundary point to the substituted integral (justified by domination with $C(1 + x^{\beta+2}) e^{-x/2}$ plus an exponentially small tail), then cancels exponentials in $Z(2\Phi_d - \Phi_q)\, Z(\Phi_q) / Z(\Phi_d)^2$ and expands $(2\Phi_d - \Phi_q)^{-\alpha} = 2^{-\alpha}\, \Phi_d^{-\alpha}(1 + \alpha\, \Phi_q / (2\Phi_d) + O(\Phi_d^{-2}))$. Under mere regular variation, Karamata's Laplace theorem replaces Watson and returns the slowly varying factor. The essential edge sets $\phi_\lambda(t) = \lambda\, t + c\, t^{-\rho}$; the strictly convex ϕ_λ has unique minimizer $t_\lambda = (c\, \rho / \lambda)^{1/(\rho+1)}$ and value $\phi_\lambda(t_\lambda) = K\, \lambda^p$, with $\phi_\lambda''(t_\lambda) = (\rho+1)\, \lambda / t_\lambda$; the one-dimensional saddle-point expansion (de Bruijn Ch. 6) then gives $\int_0^{t_0} e^{-\lambda t} f^*(t)\, dt = A_E\, \lambda^{-m} e^{-K\, \lambda^p}(1 + O(\lambda^{-\min\{p, \eta_E/(\rho+1)\}}))$. Substitution in the overlap gives the C_E prefactor and the $K(2 - 2^p)$ rate. The Gaussian branch is exact: $\log[Z(2d - q)\, Z(q) / Z(d)^2] = (s_{\nu^2}/2)((2d - q)^2 + q^2 - 2d^2) = s_{\nu^2}\, (d - q)^2$, the μ terms cancelling. The joint-scaling limit follows from Karamata's uniform-convergence theorem: $L((2 - r)d)\, L(r\, d) / L(d)^2 \to 1$ and $d^{2\alpha} / [((2 - r)\, d)^\alpha\, (r\, d)^\alpha] \to (r\, (2 - r))^{-\alpha}$, with the $c\, \Gamma(\alpha)$ prefactors cancelling completely ($c \cdot c / c^2 = 1$), leaving no $Z(\Phi_q)$-dependent factor. The proof uses BGT 1987 §4.12 for the Laplace-Tauberian form, de Bruijn 1981 Ch. 6 for saddle-point, and Watson's boundary-point lemma. ∎

Remark 5.1 (single-axis indexing and the $\beta \to \infty$ crossover). On the regular-variation axis the atom is $\beta = -1$ (constant overlap), the continuous polynomial edge is $\beta \in (-1, \infty)$ (degree $1 + \beta$), and the essential edge is the $\beta = \infty$ end. The crossover is concrete: as $\beta \to \infty$ the polynomial degree $1 + \beta \to \infty$, so the price outgrows every fixed power, which is exactly the onset of the stretched-exponential regime (C). Writing a power edge as $t^\beta = \exp(-\beta \log(1/t))$ and letting $\beta = \beta(t) \to \infty$ with $\beta(t) \log(1/t) = \psi(1/t)$ regularly varying of index ρ recovers (C) with rate $p = \rho/(\rho+1)$. The three named cases are the two endpoints and the interior of one axis.

Remark 5.2 (scope of the regime-spectrum statement). The fixed-cap continuous-polynomial constant is $c_\beta(\Phi_q) = 2^{-\alpha} e^{-\Phi_q s^*} Z(\Phi_q) / (c_0\, \Gamma(\alpha))$, carrying the exact value $Z(\Phi_q)$; the essential-edge rate is $p = \rho/(\rho+1)$; and an atom at s^* is the degenerate endpoint $\beta = -1$ rather than the continuous $\beta = 0$ edge.

Remark 5.3 (what this lemma buys and does not buy). Lemma 5.1 sharpens Theorem 3P′ achievability, not Theorem 3P's converse: it turns the abstract overlap multiplier in the self-normalized importance-sampling budget into explicit edge-regime budgets — degree $\beta + 1$ polynomial, stretched exponential of rate $\rho/(\rho+1) < 1$, or Gaussian-exponential outside the class. It does not improve the median-of-means constants, the stage-(i) hindsight factor, or the clipping/truncation convention outside the bounded class, and it leaves §6.3's measured $1 + \hat{\chi}^2$ curve untouched (those numbers are computed numerically, not from the constant). Proposition 3N remains open in the same way as before: the lemma identifies when the overlap itself grows polynomially, stretched exponentially, or Gaussian-exponentially, but it does not remove the band-localization residual $(\Phi_d - \Phi_q)\, \mathrm{osc}(u)$ or prove an absolute-constant lower bound $n \geq c_1\, \sigma^2\, \delta^{-2}\, (1 + \chi^2)$ — that residual is a property of the Le Cam two-point geometry, not of the MGF asymptotics, so the lemma is orthogonal to it. What it does give 3N is a sharp regime boundary: the necessity factor $(1 + \chi^2) / ((\Phi_d - \Phi_q)\, \mathrm{osc}(u))$ grows in the essential and Gaussian-bulk regimes and at polynomial edges with $\beta > 0$ (factor $\asymp \Phi_d^\beta$), but is flat at the box edge $\beta = 0$, matching 3N's

own statement that the layer-cake route proves no growing factor there.

Corollary 3Z (asymmetric price visibility). The certification price's overlap factor $1 + \chi^2$ is a functional of ν alone — determined by the (OC) proxy-score telemetry (the internal loop's own tilted proxy marginals, mutually absolutely continuous across tilts) in the same idealization under which Theorem 2 proves blindness: the loop's data determines the price of the channel it cannot replace. (The identification needs proxy-score-rich (OC) telemetry; a degenerate (OC) channel that does not reveal the proxy law — e.g. a constant monitor — identifies nothing, and Theorem 2's blindness still holds for it. Theorem 2 is universal over (OC); 3Z's price-visibility is scoped to ν-identifying telemetry.) (The achievability multiplier $\tilde{\sigma}^2 = \bar{\sigma}^2 + B^2$ rests on maintained bounds — welfare-conditional variance and range — which are assumptions, not telemetry.) At finite sample the visibility is one-sided: valid lower confidence bounds on the required budget are cheap and anytime-valid (directly on the budget of the certified protocol of Theorem 3P′ — the referent against which an "available budget multiple" is defined; on the information-theoretic minimum via Theorem 3P's band functional — itself a ratio of bounded tail means, lower-boundable by the same construction — or Proposition 3N under (BD)); in the sufficient direction, the only budget proven sufficient here is the a-priori worst case — with a known proxy bound $\|u\|_\infty$, $e^{O(\Phi_d \|u\|_\infty)}$ is provably sufficient but exponential exactly in the regimes at issue. Any data-adaptive upper confidence bound must carry an explicit allowance for upper-tail mass un-hit by the sample (a two-point pair on the unseen tail invalidates any upper bound omitting it), and how far such allowances can certifiably buy down the worst case is left open here. Internal data can prove an oversight budget insufficient; the only sufficiency certificate established is the a-priori worst case. Self-certification is not merely unreliable; it is structurally one-sided.

Corollary 3CC (quantitative channel collapse; un-hit-mass deficiency floor). Fix $\Phi_q < \Phi_d$, a band $t \in \mathcal{T}$ (the feasible band set of Theorem 3P, all three conditions), and the top band $S_t = \{s : s > s^* - t\}$. Let (w_0, w_1) be Theorem 3P's two-point pair — worlds differing only by the ramp h_t supported on S_t. Both worlds are realized with W deterministic given u, $W_i(a) = w_i(u(a))$ (equivalently, with any common conditional noise: $W_1 = W_0 - h_t(u(a))$ pointwise), so that $Y \mid S = s$ has the same law in both worlds off S_t and law $N(w_i(s), \sigma^2)$ in the deterministic realization — class membership specifies only the conditional mean $w(s)$, which leaves the off-band conditional Y-law free, and the prover, choosing the hard instance, pins it. Define two w-indexed experiments: $\mathcal{E}_q^m$, the protocol of m welfare queries (S_i, Y_i), $Y_i = W(a_i) + \sigma\xi_i$, at (possibly adaptively chosen) tilts $\Phi_i \le \Phi_q$; and $\mathcal{E}_d(\varepsilon)$, a protocol of queries at tilt Φ_d with budget large enough to separate the pair to total variation $\ge 1 - \varepsilon$ (such a budget exists for every $\varepsilon > 0$: a mean test at Φ_d, Appendix C). (i) Deficiency floor (proven). The two worlds' query laws are identical except on draws landing in S_t, so $TV(P_0^m, P_1^m) \le 1 - (1 - \nu_{\Phi_q}(S_t))^m$; any test of the pair from $\mathcal{E}_q^m$ errs with probability $\ge \tfrac{1}{2}(1 - \nu_{\Phi_q}(S_t))^m$, and the Le Cam deficiency of $\mathcal{E}_q^m$ against $\mathcal{E}_d(\varepsilon)$ is $\ge (1 - \nu_{\Phi_q}(S_t))^m - \varepsilon$ under the total-variation norm convention $\|P - Q\| = 2TV(P,Q)$ (halve the display under a TV-distance convention). The floor is decreasing in m — internal sampling helps — but at the rate set by the top-band mass at the query cap, $\nu_{\Phi_q}(S_t)$, which the cap keeps small exactly where deployment pressure concentrates (monotone tail fact); asymptotically it is the un-hit-mass form $e^{-m \cdot \nu_{\Phi_q}(S_t)}$. For budget scaling in the overlap factor $1 + \chi^2$, see Theorem 3P′ and Proposition 3N; no sample-complexity clause is claimed here. (ii) One-sided collapse (missing-mass undershoot). Any plug-in estimator of $Z(\Phi_d)$ built from nonnegative importance weights on draws taken at tilts $\le \Phi_q$ represents only the mass its sample hits: on the event that no draw lands in S_t — probability $\ge (1 - \nu_{\Phi_q}(S_t))^m$ — it omits the band's contribution, a fraction $\nu_{\Phi_d}(S_t)$ of the target, entirely, and undershoots by at least that mass fraction unless the un-hit complement is overestimated. No almost-sure domination and no monotonicity in m are claimed (a resample over-representing top scores can overshoot; the measured single-resample overlap curve is non-monotone in m). The observed form is the repo telemetry: deployment-tilt subsample medians sit below the full-sample value with the lower confidence bands intact (`results/onesided_table.csv`; §6.3(iv)). Remark (observed telemetry). At deployment strength the definitive deep-tilt run (n = 300,000, tilted sampling; §6.3(iv)) shows the partition-function estimate carried ~0.9999–1.0 by a single completion with effective sample size ≈ 1, so the flat overlap estimate $(1 + \hat{\chi}^2 \approx 1.027)$ is an extreme-order-statistic object whose lower confidence sequence is identically zero — reported as measurement tied to §6.3(iv), not as a theorem clause or a monotone-drift claim. (The n = 6,400 subsample read is reported only as earlier telemetry; the deep-tilt run is the circulation-facing measurement.) (Proof of (i)–(ii): Appendix C.)

Proposition 3N (overlap necessity up to band localization; dominance hypothesis named). Let $\Phi_q < \Phi_d$, take the welfare class $\mathcal{W}(B, L)$, and let the feasible band set of Theorem 3P read at 2δ — conditions (ii)–(iii) with 8δ in

place of 4δ; write $\mathcal{T}_{2\delta}$ — be nonempty. Inverting the layer-cake inequality of the 3P′ proof gives, over all top bands, $\sup_{t>0} \nu_{\Phi_d}(S_t)^2/\nu_{\Phi_q}(S_t) \geq (\chi^2(\nu_{\Phi_d}\|\nu_{\Phi_q}) + 1 - \mathrm{LR\_min}) / ((\Phi_d - \Phi_q)\cdot\mathrm{osc}(u))$, with $\mathrm{LR\_min} \leq 1$. Call (BD) the band-dominance hypothesis: the feasible-band functional of Theorem 3P satisfies $\sup_{t\in\mathcal{T}_{2\delta}} \nu_{\Phi_d}(S_{t/2})^2/\nu_{\Phi_q}(S_t) \geq \mathrm{c\_geo}^{-1} \cdot \sup_{t>0} \nu_{\Phi_d}(S_t)^2/\nu_{\Phi_q}(S_t)$ for a constant $\mathrm{c\_geo} \geq 1$ — an explicit geometry hypothesis, not a consequence of the standing assumptions (it bridges both the feasible-set restriction and the half-band numerator); it is satisfied at atomic-top geometries for δ below an instance threshold, including the measured §6.3 instance at its calibrated parameters ($\delta = 0.08$, $\hat{L} = 1.336$, $\hat{B} = 3.280$): there the top-atom (all-band) ratio $\sup_{t>0} \nu_{\Phi_d}(S_t)^2/\nu_{\Phi_q}(S_t)$ is 3.95 and the feasible-band functional is 2.90, so (BD) holds with $\mathrm{c\_geo} \approx 1.36$ (the feasible functional equals the all-band 3.95 only for δ below ≈ 0.029 at the 2δ-read band set, ≈ 0.057 at Theorem 3P's own; both functionals are recorded in `results/band_dominance.json`). c_geo is an instance-dependent geometry constant, not an absolute prefactor: ≈ 1.36 here, but it grows on thin/high-β edges (≈ 14.8 on the log-concave instance of the delimitation below — bounded in $\Phi_d\cdot\mathrm{osc}(u)$, unbounded in edge thinness), which is exactly why $(144\cdot\mathrm{c\_geo})^{-1}$ confers no absolute constant and absolute-constant necessity stays open. Under (BD), Theorem 3P instantiated at 2δ — its pair then has $G_0 = 0$ and $G_1 \geq 2\delta$, which any test certifying $G_w(\Phi_d) \leq \delta$ versus $G_w(\Phi_d) \geq 2\delta$ with both errors ≤ 1/3 must separate — yields the proven lower bound on the minimal query budget n for that certification task: $n \geq (144\cdot\mathrm{c\_geo})^{-1}\, \sigma^2\, \delta^{-2} \cdot (1 + \chi^2(\nu_{\Phi_d}\|\nu_{\Phi_q}) - \mathrm{LR\_min}) / ((\Phi_d - \Phi_q)\cdot\mathrm{osc}(u))$, i.e. the deployment-pressure dependence through the overlap is necessary up to the band-localization residual $(\Phi_d - \Phi_q)\cdot\mathrm{osc}(u)$, which divides the lower bound — the same residual appearing in Theorem 3P″'s current lower/upper comparison. The certified importance-sampling protocol of Theorem 3P′ attains $n \leq c_2\, \tilde{\sigma}^2\, \delta^{-2} \cdot (1 + \chi^2(\nu_{\Phi_d}\|\nu_{\Phi_q})) \cdot \log(N_{grid}/\alpha)$, where c_2 carries Theorem 3P″'s additive stage-(i) hindsight factor — $c_2 = C\cdot(2 + \min(\Phi_q\cdot\mathrm{osc}(u), K_g(\Phi_q)))$ — so c_2 is Φ_d-free but not osc-free (it depends on $\Phi_q\cdot\mathrm{osc}(u)$; that dependence is part of what the open absolute-constant question below is about). Two honest delimitations. First, absolute-constant necessity — $n \geq c_1\sigma^2\delta^{-2}(1 + \chi^2)$ with c_1 free of $(\Phi_d, \mathrm{osc}(u))$ — is open: the layer-cake route provably cannot deliver it (a log-concave instance has $1 + \chi^2 = 3.01$ while the all-band sup over $t \in (0, \mathrm{osc}(u)]$ (proper bands; every feasible $\mathcal{T}$ lies there, since condition (i) forces $\nu(S_t) \leq ¼$) of Theorem 3P's half-band-numerator functional $\nu_{\Phi_d}(S_{t/2})^2/\nu_{\Phi_q}(S_t)$ is 0.068 (the S_t-numerator layer-cake functional sups at 1.0), a shortfall growing linearly in $\Phi_d\cdot\mathrm{osc}(u)$; Appendix C). Second, the residual-divided bound certifies growing necessity only where $(1 + \chi^2)/((\Phi_d - \Phi_q)\cdot\mathrm{osc}(u))$ grows — polynomial edges with $\beta > 0$ (factor $\asymp \Phi_d{}^\beta$), the essential edge, and faster regimes (equivalently wherever $1 + \chi^2$ grows superlinearly in Φ_d; Lemma 5.1, Remark 5.3); at a polynomial edge with $\beta = 0$ this route proves no growing factor. What survives: on (BD) geometries no (OC)-internal protocol at query cap Φ_q can collapse the overlap factor below the residual-divided floor without a welfare-touching channel; off (BD) geometries, Theorem 3P's feasible-band functional remains the proven floor. (Proof: Appendix C.)

Definition (ρ-grader). For $\rho \in [0,1]$ and noise scale $\sigma_0 > 0$, a ρ-grader is a per-query channel that, on a queried action a drawn at tilt $\Phi \leq \Phi_q$, returns the verdict $V = \rho\cdot W(a) + \sqrt{(1-\rho^2)}\cdot\sigma_0\cdot\xi$, $\xi \sim N(0,1)$ independent of everything else, possibly alongside arbitrary (OC)-measurable side information. This is a modeling definition of a channel class, not a derived fact. The Gaussian member is canonical; the class extends to (a) any channel whose verdict, conditionally on the trajectory, is Gaussian with welfare-independent (trajectory-measurable) variance $v(s) \geq (1-\rho^2)\sigma_0{}^2$ and mean separation at most $\rho\cdot|\Delta w|$ under a welfare perturbation Δw (the Gaussian KL formula then gives the same per-query bound), and (b) any post-processing of the canonical read (data processing only lowers KL). Welfare independence of the variance in (a) is required, not decorative: a welfare-dependent variance is itself a verdict-channel side-band whose KL is not controlled by the mean-shift formula, exactly parallel to the sub-Gaussian exclusion below. Honest delimitation of the class: "attenuated mean plus sub-Gaussian noise" alone is not admissible in a converse — a lower bound needs an upper bound on per-query KL, and shifted non-Gaussian noise at the same sub-Gaussian parameter can have arbitrarily large, even infinite, KL (bounded-support noise under a mean shift); the class is therefore defined by Gaussianity-or-degradation, not by moment conditions. Endpoints: $\rho = 0$ makes the verdict welfare-free, so the channel satisfies (OC) and Theorem 2 applies; $\rho = 1$ is the noiseless gold read ($\sigma_{eff} = 0$ below — §5's query model in its $\sigma \to 0$ limit, not its σ-noise form); §5's query model at noise σ is the known-ρ rescaling of any member on the matched locus $\sigma_{eff}(\rho, \sigma_0) := \sqrt{(1-\rho^2)}\cdot\sigma_0/\rho = \sigma$ (e.g. $\rho = 1/\sqrt{2}$, $\sigma_0 = \sigma$).

Proposition 3G (graded oversight: the ρ-interpolation). Fix $\Phi_q < \Phi_d$, the welfare class $\mathcal{W}(B, L)$ and feasible band set $\mathcal{T}$ of Theorem 3P, and a ρ-grader with $\rho \in (0,1)$, $\sigma_0 > 0$; write $\sigma_{eff}^2(\rho) = (1-\rho^2)\sigma_0{}^2/\rho^2 \in (0,\infty)$. (i) Lower

bound (proven). Theorem 3P holds verbatim with σ^2 replaced by $\sigma\_\mathrm{eff}^2(\rho)$: any test of $G = 0$ versus $G \geq \delta$ with both errors $\leq 1/3$ from n grader queries at tilts $\leq \Phi\_q$ (adaptive allowed, arbitrary (OC) side stream free) requires $n \geq (1/36) \cdot ((1-\rho^2)/\rho^2) \cdot {\sigma_0}^2 \cdot \delta^{-2} \cdot \sup\_{t\in\mathcal{T}} \nu\_{\Phi\_d}(S\_{t/2})^2/\nu\_{\Phi\_q}(S\_t)$. The budget diverges as $\rho \to 0$ — Theorem 2's blindness recovered continuously, with exact (OC) blindness at $\rho = 0$ — and reproduces Theorem 3P exactly on the matched locus $\sigma\_\mathrm{eff} = \sigma$. Degenerate endpoint: at $\rho = 1$, $\sigma\_\mathrm{eff}^2 = 0$ (noiseless gold reads) and the display is vacuous, exactly as Theorem 3P's is in its $\sigma^2 \to 0$ limit; the floor that survives there is Corollary 3CC's un-hit-mass deficiency bound, which is noise-free and holds for every $\rho \in [0,1]$, since the two worlds' verdict laws coincide off the band at any ρ. (ii) Achievability, ρ known (proven). If ρ is known, the rescaled read $V/\rho = W(a) + \sigma\_\mathrm{eff}(\rho)\cdot\xi$ is exactly §5's query channel at noise $\sigma\_\mathrm{eff}(\rho)$, so Theorem 3P′ applies verbatim with $\tilde{\sigma}\_\mathrm{eff}^2 := (1-\rho^2){\sigma_0}^2/\rho^2 + \sup\_s \mathrm{Var}(W|u{=}s) + B^2$ — §5's achievability convention $\tilde{\sigma}^2 = \bar{\sigma}^2 + B^2$ with channel noise $\sigma^2 \to \sigma\_\mathrm{eff}^2(\rho)$ — and the 3P/3P′ matched orders, and Proposition 3N's necessity statement under (BD), are inherited wholesale with $(\sigma^2, \tilde{\sigma}^2) \to (\sigma\_\mathrm{eff}^2, \tilde{\sigma}\_\mathrm{eff}^2)$. This requires ρ known. With only a certified lower bound $\rho \geq \rho\_\mathrm{min} > 0$, rescaling by ρ_min estimates the inflated gap $(\rho/\rho\_\mathrm{min})\cdot G\_w \geq G\_w$, so a passed certificate remains valid — certification stays sound, one-sidedly conservative — at the ρ_min price: all three $\tilde{\sigma}^2$ terms inflate by at most ρ_min^{-2} (and B enters N_grid only logarithmically). Without any certified ρ_min, no certification is conferred (next remark). (Proof: Appendix C.)

Remark (the grader's credentials sit one level up). The grader's own informativeness ρ is not identifiable from any data whose conditional law given the trajectory is fixed, because the verdict's law differs between welfare worlds only through W. Exhibited pair: scenario A — Appendix B's world w_0 ($\Delta = 0$, W deterministic given u) overseen by a genuine ρ-grader; scenario B — world w_1 (Δ large) overseen by an (OC) simulator returning $N(\rho\cdot w_0(u(a)), (1-\rho^2){\sigma_0}^2)$, a fixed trajectory-measurable kernel that never reads W. The trajectory laws coincide across the two worlds (Theorem 2) and the verdict kernels are identical by construction, so A and B generate identical observation laws at every sample size: no test computed from trajectory-plus-verdict data distinguishes ⟨informative grader, safe world⟩ from ⟨welfare-blind grader, unsafe world⟩, and any certificate of $\rho \geq \rho\_\mathrm{min} > 0$ computed from such data would do exactly that. Certifying the grader's informativeness is therefore itself a welfare-touching measurement — gold audits of the grader's verdicts against realized welfare — and is priced by this section one level up: Theorem 2 applies to the grader's credentials, and Corollary 3Z's one-sidedness extends to them. Consequence for practice: an unverified grader confers no certification, whatever its verdicts say. Scope, stated precisely: this is non-identifiability by an exhibited construction (the deterministic-given-u realization with the canonical Gaussian grader), which is all a non-identifiability claim requires; no claim is made beyond the exhibited pair.

Remark (debate, delimited and priced within the ρ-grader class). Debate-style oversight (Brown-Cohen et al.) escapes the impossibility exactly through whatever grounding makes the judge's verdict law depend on W — code that runs, experiments, outcome checks. Grounded judging is $\rho > 0$ in the ρ-grader model, and grounded judging whose verdict channel lies in the ρ-grader class is priced by (i)–(ii); grounding outside the Gaussian-or-degraded class (e.g. a noiseless binary outcome check, which is no post-processing of any canonical read) may be cheaper to inform — the Definition's exclusion — and the $(1-\rho^2)/\rho^2$ multiplier is not claimed for it; what it retains is Corollary 3CC's un-hit-mass floor, which is channel-shape-free. Transcript-only judging is $\rho = 0$ and Theorem 2 applies (§7).

Appendix A: Proofs for Section 3

Lemma 1. μ_0 finite and $|u| \leq \|u\|\infty$ give $|e^{zu}| \leq e^{|\mathrm{Re}\ z|\|u\|\infty} \in L^1(\mu_0)$; $Z(z) = \int e^{zu}d\mu_0$ and $N(z) = \int We^{zu}d\mu_0$ are entire (Morera + Fubini on triangles); $Z > 0$ on $\mathbb{R}$, so $T\_w = N/Z$ is real-analytic on $\mathbb{R}$. Differentiation under the integral (dominated, locally uniform in Φ): $T\_w' = (N'Z - NZ')/Z^2 = E\_\Phi[Wu] - E\_\Phi[W]E\_\Phi[u] = \mathrm{Cov}\_{\pi\_\Phi}(u, W)$. ∎

Lemma 2. $e^{\Phi u(a)}$ is $\sigma(u)$-measurable; by the tower property $E\_{\pi\_\Phi}[W] = E\_{\mu_0}[W\ e^{\Phi u}]/Z(\Phi) = E\_{\mu_0}[E[W|u]\ e^{\Phi u}]/Z(\Phi) = \int w(s)e^{\{\Phi s\}\nu(ds)/\int e^{\{\Phi s\}\nu(ds)}$. ∎

Theorem 1. (a)/(b): under (G3) the limit L exists; T_w is continuous and $\to L$, so if $\Delta := \sup\_{\{\Phi\geq 0\}} T\_w - L > 0$, the supremum exceeds the tail values eventually and is attained on a compact interval, hence at some $\Phi^* < \infty$; G2 gives $T\_w'(0) = \mathrm{Cov}\_{\mu_0}(u,W) > 0$, so $\Phi^* > 0$ is interior; at an interior maximum of a non-constant real-analytic function, $T\_w'(\Phi^*) = 0$, i.e. $\mathrm{Cov}\_{\pi\_{\Phi^*}}(u,W) = 0$; "never exceeds the peak" is the definition of

the (attained) supremum; the peak-to-limit decline is Δ by definition. (c): $T\_w'(0) < 0$ and continuity give decline below baseline on an initial interval; global monotone decline is claimed only per instance (§6.1: verified by quadrature, App. D; an analytic route is positivity of $\mathrm{Cov}_{\pi_\Phi}(u, \mu)$ along the truncated-Gaussian path). Laplace concentration for (G3)-sufficiency: $\pi_\Phi(u \geq s^* - \delta) \to 1$ for all $\delta > 0$ since $\nu_\Phi(u < s^*-\delta)/\nu_\Phi(u \geq s^*-\delta/2) \leq e^{-\Phi\delta/2}\cdot\nu(u<s^*-\delta)/\nu(u\geq s^*-\delta/2) \to 0$; if $w(s) \to w(s^*-)$ as $s \uparrow s^*$, then $T\_w(\Phi) \to w(s^*-)$; if $\nu(\{s^*\}) > 0$, $T\_w(\Phi) \to w(s^*)$. Scope notes: the $W = u + \varepsilon v$ perturbation claim requires $v \in C^2$ near an interior nondegenerate maximum of u that is the unique global maximizer (μ_0-essentially, with Laplace localization: sup of u outside every neighborhood strictly below u^*; finite-dimensional chart) and μ_0 with locally C^1 (or Lipschitz) positive density at the maximizer; under those hypotheses one chart controls all the mass, the second-order Laplace expansion $\mathrm{Var}_\Phi(u) \asymp c\_u/\Phi^2$, $\mathrm{Cov}_\Phi(u,v) = O(1/\Phi^2)$ is standard, giving $T' > 0$ beyond an instance scale Φ_0 when $\varepsilon < c\_u/\sup|c\_v|$, and on $[0, \Phi_0]$ by compactness ($\mathrm{Var}_\Phi(u)$ continuous and positive), hence $\Delta = 0$ for all ε below an instance constant; boundary maxima, densities thin or vanishing at the maximizer, and proxies with multiple global maximizers (ties at the top) are outside the claim. The (G3)-necessity counterexample, as a complete instance: $\mu_0 =$ uniform on $[-1,1]$ (or any compactly supported law with density continuous and positive at 0), $u = -x^2$, $W = \sin(\log(1/|x|))$; the trajectory oscillates forever with exact asymptotic amplitude $0.631 = |\Gamma((1-i)/2)|/\sqrt{\pi} = 1/\sqrt{\cosh(\pi/2)}$. ∎

Appendix B: Proof of Theorem 2

Fix (μ_0, u) and a channel satisfying (OC): conditional on the action trajectory, M is drawn from a kernel fixed across welfare functions (the u-process plus independent-noise model is only a special case). The trajectory $(\pi\_\Phi)\Phi$ is a functional of (μ_0, u) alone; hence the joint law of (trajectory, M) is a functional of the trajectory law and the fixed kernel, and is invariant to w. Construction: let w_0 have $\Delta_0 = 0$ ($w_0 \equiv 0$ suffices; any nondecreasing w_0 with a limit at s^* works); let h be nondecreasing with $0 \leq h \leq \delta\_\mathrm{target} \leq B$, supported on $(s^*-t, s^*]$, continuous at s^* with $h(s^*-) = \delta\_\mathrm{target}$; set $w_1 = w_0 - h$. Both define valid welfare functions (any W with the prescribed conditional mean and bounded range realizes them). Observation laws: identical, since neither the trajectory law nor the channel kernel changed. Trajectories: $T\{w_1\}(\Phi) = T\_\{w_0\}(\Phi) - \int h\, d\nu\_\Phi$; the subtracted term $\to h(s^*-) = \delta\_\mathrm{target}$ (Laplace, as in App. A) and is at most $\delta\_\mathrm{target}\cdot\nu(S\_t)$ at $\Phi = 0$. Thus, whenever top bands can be chosen with small baseline mass, Δ differs between worlds — by an amount growing with δ_target within the bound B — and with it the pair (Δ, Φ^*). Degenerate proxy laws already concentrated at the top do not support the large-gap clause; the W-invariance of the observation law still holds. The construction does not exhibit two worlds with distinct finite interior Φ^*. Identical laws are indistinguishable by any measurable rule, deterministic or randomized, at any sample size: error probability is exactly $1/2$ against the uniform prior over the pair. Remark 2 (interpretability blindness): the policy weights realize the same stochastic process in both worlds because the optimizer's update rule is (OC)-measurable; the two worlds differ in W, not in the system. ∎

Appendix C: Proofs for Section 5 (channel pricing)

Monotone tail fact. For $\Phi' > \Phi$, $d\nu\_\{\Phi'\}/d\nu\_\Phi \propto e^{(\Phi'-\Phi)s}$ is increasing $\Longrightarrow$ likelihood-ratio order $\Longrightarrow$ $\nu\_\Phi(S)$ nondecreasing in Φ for upper sets S and $E\_\{\nu\_\Phi\}[g]$ nondecreasing for nondecreasing g. ∎

Theorem 3R. Grid $\Phi\_j = j\Phi\_d/N\_\mathrm{grid}$, $N\_\mathrm{grid} = \lceil 4\Phi\_d\|u\|_\infty B/\delta\rceil$, from the modulus $|T\_w'| = |\mathrm{Cov}\{\pi\_\Phi\}(u,W)| \leq \|u\|_\infty B$ (Cauchy–Schwarz). Batches at tilts spaced $1/\mathrm{osc}(u)$ apart; a batch at Φ_j serves grid tilts with $|\Phi-\Phi\_j| \leq 1/\mathrm{osc}(u)$ by importance reweighting with weight second moment $\leq \exp(2\sup\_\Phi \mathrm{Var}\{\nu\_\Phi\}(s)\cdot\Delta\Phi^2) \leq e^{1/2}$ on the window ($\mathrm{Var}\_\{\nu\_\Phi\}(s) \leq \mathrm{osc}(u)^2/4$). Each batch of size $n/(\lceil\Phi\_d\cdot\mathrm{osc}\rceil+1)$ estimates all its windowed grid values within $\delta/8$ w.p. $1-\alpha/N\_\mathrm{grid}$ by sub-Gaussian concentration (variance $\tilde\sigma^2 = \bar\sigma^2 + B^2$, bounded weights on-window); union bound; reject when $\max\_j \hat T(\Phi\_j) - \hat T(\Phi\_d) > 3\delta/8$ (H_0 statistic $\leq \delta/4 < 3\delta/8 < \delta/2 \leq H_1$). Total $n \leq C\tilde\sigma^2\delta^{-2}(1+\Phi\_d\cdot\mathrm{osc}(u))\log(N/\alpha)$. Variance-adaptive schedule: replace equispaced batches by the greedy exact-χ^2 cover — from a window's left edge a, extend to the largest r with $Z(2r-a)Z(a)/Z(r)^2 \leq e^{1/2}$; every other step is unchanged with one batch per greedy window, giving the $\min(\Phi\_d\cdot\mathrm{osc}(u), K\_g)$ factor of the statement. The count obeys $K\_g \asymp \int\_0^{\Phi\_d}\sqrt{(2\cdot\mathrm{Var}\_\{\nu\_\Phi\}(s))}d\Phi$ up to constants (order-equivalence per the companion needle note, numerically verified, not proven exact), and the integral is $\leq \sqrt{(2\Phi\_d\cdot\mathrm{osc}(u))}$ by Cauchy–Schwarz with the cumulant identity $\int\_0^{\Phi\_d}\mathrm{Var}\_\{\nu\_\Phi\}(s)d\Phi = m(\Phi\_d) - m(0) \leq \mathrm{osc}(u)$, $m = (\log Z)'$; the equispaced count is the Popoviciu worst case ($\mathrm{Var} \leq \mathrm{osc}^2/4$ substituted inside the schedule). The window rule must use the exact

within-window second moment (equivalently for the e^{1/2} guarantee, sup-of-Var over the window), not the left-endpoint variance, which violates the e^{1/2} bound by 10× on bimodal bases. ∎

Theorem 3P. Worlds $w_0 \equiv 0$ and $w_1 = -h_t$, $h_t(s) = \min\{h, L(s-(s^*-t))_+\}$, $h \le \min(Lt, B)$. Both worlds are realized with W deterministic given u, $W_i(a) = w_i(u(a))$ (equivalently, with any common conditional noise: $W_1 = W_0 - h_t(u(a))$ pointwise), so $Y \mid S = s$ has law $N(w_i(s), \sigma^2)$ in the deterministic realization and the off-band conditional Y-laws coincide; mean-only specification would not fix the off-band Y-laws (worlds with equal off-band means but different conditional W-laws have different Y-laws there), and the prover may choose the realization. Gap: since h_t is nondecreasing, $w_1 = -h_t$ has $T_{w_1}(\Phi)$ nonincreasing by the monotone tail fact, so $G_1(\Phi_d) = \int h_t \, d\nu_{\Phi_d} - \int h_t \, d\nu \ge (h/2)\nu_{\Phi_d}(S_{t/2}) - h\nu(S_t) \ge (h/4)\nu_{\Phi_d}(S_{t/2})$ under feasibility condition (i) $\nu(S_t) \le \tfrac{1}{4}\nu_{\Phi_d}(S_{t/2})$; with $h = 4\delta/\nu_{\Phi_d}(S_{t/2})$ (admissible under conditions (ii) $Lt \ge 4\delta/\nu_{\Phi_d}(S_{t/2})$ and (iii) $\nu_{\Phi_d}(S_{t/2}) \ge 4\delta/B$), $G_1(\Phi_d) \ge \delta$. KL per query at tilt $\Phi \le \Phi_q$: s-marginals are world-independent; the Y-laws are $N(0,\sigma^2)$ vs $N(-h_t(s),\sigma^2)$, so $KL = E_{\nu_\Phi}[h_t^2]/(2\sigma^2) \le h^2\nu_{\Phi_q}(S_t)/(2\sigma^2)$ by the tail fact. Adaptivity: divergence decomposition — $KL(P_0^n \| P_1^n) = \Sigma_i \, E_0[KL(P_0(\cdot|\mathcal{F}_{i-1}) \| P_1(\cdot|\mathcal{F}_{i-1}))] \le n \cdot \max_{\Phi \le \Phi_q}$ per-query KL; the free (OC) stream is law-identical (App. B) and contributes zero, including to adaptive tilt choices. Le Cam: both errors $\le 1/3 \Longrightarrow$ TV $\ge 1/3 \Longrightarrow$ (Pinsker) KL $\ge 2/9 \Longrightarrow n \ge (1/36)\sigma^2\delta^{-2}\nu_{\Phi_d}(S_{t/2})^2/\nu_{\Phi_q}(S_t)$; take sup over the feasible set $\mathcal{T}$. Nonemptiness of $\mathcal{T}$ is an instance hypothesis, not a consequence of the standing assumptions: at an atomic or concentrated top with $p_d := \lim_{t\to 0} \nu_{\Phi_d}(S_{t/2}) > 0$ and $\nu(S_t) \le p_d/4$, conditions (i)–(iii) hold on $t \in [4\delta/(Lp_d), t_0]$ whenever $\delta \le (p_d/4)\cdot\min(Lt_0, B)$ and $p_d \ge 4\delta/B$, so $\mathcal{T} \supseteq [4\delta/(Lp_d), t_0] \ne \varnothing$; on a thin no-atom edge ($\nu_{\Phi_d}(S_{t/2}) \to 0$ as $t \to 0$) condition (i) can fail for every t and $\mathcal{T} = \varnothing$ for all δ, where the displayed sup is over the empty set and a separate instance check is required (the §6.3 atomic instance satisfies the sufficient condition). Branch remark: a common nondecreasing component g added to both worlds leaves the (OC) law and the per-query KL unchanged (the worlds differ by the same h_t); the bound concerns the gap functional, not the branch classification. ∎

Theorem 3P′. Stage (i) is Theorem 3R on $[0, \Phi_q]$.

Stage (ii) handles $\Phi \in (\Phi_q, \Phi_d]$. Define

$$r_\Phi(s) = e^{(\Phi-\Phi_q)s} Z(\Phi_q)/Z(\Phi).$$

Then

$$T_w(\Phi) = \frac{E_{\nu_{\Phi_q}}[w(s) r_\Phi(s)]}{E_{\nu_{\Phi_q}}[r_\Phi(s)]}.$$

The weights are bounded by $e^{(\Phi-\Phi_q)s^*} Z(\Phi_q)/Z(\Phi)$, and their second moment is

$$E_{\nu_{\Phi_q}}[r_\Phi^2] = \frac{Z(2\Phi - \Phi_q) Z(\Phi_q)}{Z(\Phi)^2}.$$

This second moment is nondecreasing in Φ on $(\Phi_q, \Phi_d]$, since

$$\frac{d}{d\Phi}\{\log Z(2\Phi - \Phi_q) - 2\log Z(\Phi)\} = 2[m(2\Phi - \Phi_q) - m(\Phi)] \ge 0,$$

where $m = (\log Z)'$ is nondecreasing and $m' = \mathrm{Var}_{\nu_\Phi}(s) \ge 0$. Hence the maximum occurs at Φ_d, with value $1 + \chi^2(\nu_{\Phi_d} \| \nu_{\Phi_q})$.

Stage-(ii) concentration, with the bounded-class achievability constants in full. Fix a grid tilt $\Phi \in (\Phi_q, \Phi_d]$. On the standing bounded class the weights are bounded a.s. by $R := e^{(\Phi-\Phi_q)s^*} Z(\Phi_q)/Z(\Phi)$, with $E_{\nu_{\Phi_q}}[r_\Phi] = 1$ and $E_{\nu_{\Phi_q}}[r_\Phi^2] = 1 + \chi^2(\nu_\Phi \| \nu_{\Phi_q}) \le 1 + \chi^2$. The self-normalized estimate is

$\hat{T} = \hat{N}/\hat{D}$, with $\hat{N}$ the estimate of E_{ν_{Φ_q}}[Y r_Φ] = T_w (E_{ν_{Φ_q}}[r_Φ] = 1, E[Y|S=s] = w(s)) and $\hat{D}$ of E_{ν_{Φ_q}}[r_Φ] = 1. The identity

$$\hat{T} - T_w = \frac{\hat{N} - T_w}{\hat{D}} - \frac{T_w(\hat{D} - 1)}{\hat{D}}$$

gives, on the event $\hat{D} \geq ½$, $|\hat{T} - T\_w| \leq 2(|\hat{N} - T\_w| + B|\hat{D} - 1|)$ because $|T\_w| \leq B$. Estimate $\hat{N}$ and $\hat{D}$ by median-of-means over $K = \lceil 8 \log(4N\_grid/\alpha) \rceil$ batches each. The per-sample numerator Y r_Φ has second moment $\leq (\tilde{\sigma}^2 + B^2) \cdot E_{\{\nu_{\{\Phi\_q\}}\}}[r\_\Phi^2] = \tilde{\sigma}^2(1 + \chi^2)$ (Y = w(s) + noise, $|w| \leq B$); the per-sample denominator r_Φ has variance $\chi^2 \leq 1 + \chi^2$ and range R. Median-of-means at batch size $O(\tilde{\sigma}^2(1 + \chi^2)\delta^{-2})$ then yields $|\hat{N} - T\_w| \leq \delta/36$ and $|\hat{D} - 1| \leq \delta/(36B)$ (hence $\hat{D} \geq ½$ once $\delta \leq 18B$), each at a single grid tilt w.p. $\geq 1 - \alpha/(2N\_grid)$ — R enters the Bernstein bound only logarithmically and is dominated.

Union over the N_grid grid tilts and the two estimators costs α/N_grid per point, so $|\hat{T} - T\_w| \leq 2(\delta/36 + B \cdot \delta/(36B)) = \delta/9$ at every grid tilt with $n\_2 \leq C\tilde{\sigma}^2\delta^{-2}(1 + \chi^2)\log(N\_grid/\alpha)$. On the bounded class $r\_\Phi \leq R$ already, so no clipping is needed and the bias is zero. Thus stage (ii) gives deviation $\delta/9$ per grid point w.p. $1-\alpha/N\_grid$ with $n\_2 \leq C\tilde{\sigma}^2\delta^{-2}(1+\chi^2)\log(N\_grid/\alpha)$; the χ^2 closed form is

$$\int \left(\frac{d\nu_{\Phi_d}}{d\nu_{\Phi_q}} \right)^2 d\nu_{\Phi_q} = \frac{Z(2\Phi_d - \Phi_q)Z(\Phi_q)}{Z(\Phi_d)^2}.$$

Certification of $G \leq \delta$ vs $\geq 2\delta$ follows since $|\hat{G} - G| \leq \delta/4$ (grid modulus) + $\delta/8$ (worst-case stage-(i) estimate, at Theorem 3R's accuracy) + $\delta/9$ (stage-(ii) estimate) = $35\delta/72 < \delta/2$. Unbounded clipping is not part of this certified constant stack; the Gaussian-bulk idealization remains an asymptotic comparison outside the standing bounded theorem until a separate clipping budget is supplied. Matching: $\nu_{\{\Phi\_d\}}(S)^2 \leq \nu_{\{\Phi\_q\}}(S) \cdot E_{\{\nu_{\{\Phi\_q\}}\}}[r^2 1\_S] \leq \nu_{\{\Phi\_q\}}(S)(1+\chi^2)$ (Cauchy–Schwarz); conversely the layer-cake/level-set bound gives $1+\chi^2 \leq LR\_min + \sup_t[\nu_{\{\Phi\_d\}}(S\_t)^2/\nu_{\{\Phi\_q\}}(S\_t)] \cdot \log(LR\_max/LR\_min)$ with $\log(LR\_max/LR\_min) = O((\Phi\_d - \Phi\_q)osc(u))$ since LR level sets are top bands. ∎

Corollary 3Z. (i) The overlap factor $1 + \chi^2$ is a functional of ν alone; ν is determined by the (OC) proxy-score law across tilts (mutual absolute continuity of tilts) — i.e. by telemetry that reveals the proxy marginals, not by an arbitrary (OC) channel that may reveal nothing. (ii) Estimating $Z(2\Phi\_d - \Phi\_q)$ by importance sampling from the deepest realized tilt Φ_d has per-sample relative second moment $1 + \chi^2(\nu_{\{2\Phi\_d - \Phi\_q\}} \| \nu_{\{\Phi\_d\}})$ — the same overlap functional at the same gap. (iii) Lower confidence bounds: for queries drawn i.i.d. at a fixed realized tilt Φ_r ($\Phi\_r = 0$ gives $Z(\Phi)$ itself), the ratio $Z(\Phi)/Z(\Phi\_r) = E_{\{\nu_{\{\Phi\_r\}}\}}[e^{(\Phi - \Phi\_r)u}]$ is, for $\Phi > \Phi\_r$, the mean of a nonnegative variable bounded by $e^{(\Phi - \Phi\_r)\bar{s}}$ for any known upper bound $\bar{s} \geq s^*$ (the standing bounded class supplies $\|u\|\infty$), so a betting / empirical-Bernstein test supermartingale on the [0,1]-rescaled variable (Waudby-Smith–Ramdas) gives, by Ville's inequality, an anytime-valid lower confidence sequence on the ratio, hence on $Z(\Phi)$ when $\Phi\_r = 0$ — nonnegativity of the summands alone does not, the bounded-mean supermartingale construction does; combining the factor sequences costs one α-split. To lower-bound the required overlap itself, use the realized-tilt product decomposition: with $\Delta := \Phi\_d - \Phi\_q > 0$, $\chi^2 + 1 = E\{\nu_{\{\Phi\_d\}}[e^{\{\Delta u\}}] \cdot E_{\{\nu_{\{\Phi\_d\}}\}}[e^{\{-\Delta u\}}]$, a product of two means under the realized deployment tilt (direct computation: $E_{\{\nu_{\{\Phi\_d\}}\}}[e^{\{\Delta u\}}] = Z(2\Phi\_d - \Phi\_q)/Z(\Phi\_d)$ and $E_{\{\nu_{\{\Phi\_d\}}\}}[e^{\{-\Delta u\}}] = Z(\Phi\_q)/Z(\Phi\_d)$). The proposed certificate takes betting / empirical-Bernstein lower confidence sequences at level $\alpha/2$ on each factor (cap $e^{\{\Delta u\}}$ at any constant for the capped-mean lower confidence bound — no known s^* required; $e^{\{-\Delta u\}} \leq 1$ when $u \geq 0$) and multiplies them. On the witness instance it recovers $\chi^2 + 1 = 1.97$, and un-hit upper-tail mass now manifests as conservative silence (a lower confidence bound near the trivial floor), not invalidity — consistent with the §6.3(iv) degeneracy story. Scope, stated honestly: this proposed certificate is for budget insufficiency in pricing the realized run only — deployment-tilt samples exist by definition of the loop's own telemetry, matching this corollary's ν-identifying scoping. The threshold an "available budget multiple" is measured against is the certified SNIS estimator-family budget of Theorem 3P′, $n \leq c_2\tilde{\sigma}^2\delta^{-2}(1+\chi^2)\log(N\_grid/\alpha)$, evaluated at the lower-confidence overlap (the provisioned n divided by that display is the multiple); for protocol-free insufficiency, compare instead against Proposition 3N's converse divided by $c\_geo \cdot (1+(\Phi\_d - \Phi\_q) \cdot osc(u))$. Prospective pricing

of a not-yet-realized deeper tilt from capped-tilt batches has no exhibited non-vacuous certificate and remains open. Remark (scope of the realized-tilt certificate). The deterministic-denominator lower-bound route is valid but vacuous as exhibited: since the pointwise upper bound on Z(Φ_d) caps the lower bound on χ^2+1 at 1, it certifies nothing. The realized-tilt product certificate has been checked on a witness instance and should be treated as provisional until human mathematical referee sign-off. Upper direction: with $\|u\|\infty$ known, Z(2Φ_d−Φ_q) ≤ e^{(2Φ_d−Φ_q)‖u‖∞} and Z(Φ_d) ≥ e^{Φ_d·m(0)} give the a-priori bound $\chi^2+1 \le$ e^{O(Φ_d‖u‖_∞)}, provably sufficient when provisioned; a two-point pair on the unseen tail (within the known bound) breaks any claimed data-driven upper confidence bound that omits an explicit allowance for un-hit upper-tail mass, and the rate at which such allowances can certifiably improve on the a-priori worst case is left open here. ∎

Corollary 3CC. (i) Fix t ∈ $\mathcal{T}$ and Theorem 3P's pair $w_0 \equiv 0$, $w_1 = -h_t$, with h_t(s) = min{h, L(s−(s*−t))$_+$} supported on S_t = {s : s > s*−t}, realized as in Theorem 3P's proof with W deterministic given u, W_i(a) = w_i(u(a)) (equivalently, with any common conditional noise: $W_1 = W_0 - h_t(u(a))$ pointwise) — the pinning is needed because mean-only specification leaves the off-band conditional Y-laws free. In $\mathcal{E}$_q^m the i-th query at tilt Φ_i ≤ Φ_q returns (S_i, Y_i) with Y_i | S_i = s ~ N(w_i(s), σ^2) in the deterministic realization; the s-marginals are world-independent, and off S_t the conditional Y-laws coincide (h_t ≡ 0 there; under the pointwise coupling they coincide for any common conditional noise), so the sequential likelihoods factor identically on the event A_m = {no S_i ∈ S_t}. Each query hits S_t with conditional probability ν_{Φ_i}(S_t) ≤ ν_{Φ_q}(S_t) (monotone tail fact, applied conditionally on the past — adaptive tilt choices are covered), so P_w(A_m) ≥ $(1 - \nu_{\Phi_q}(S_t))^m$ in both worlds. Since P_0^m and P_1^m coincide on A_m, TV(P_0^m, P_1^m) ≤ max_w P_w(A_m^c) ≤ $1 - (1 - \nu_{\Phi_q}(S_t))^m$. Any test's total error is ≥ 1 − TV ≥ $(1 - \nu_{\Phi_q}(S_t))^m$, i.e. error probability ≥ ½$(1 - \nu_{\Phi_q}(S_t))^m$ against the uniform prior. $\mathcal{E}$_d(ε) exists: at tilt Φ_d the worlds' means differ by $\int$h_t dν_{Φ_d} ≥ (h/2)ν_{Φ_d}(S_{t/2}) ≥ 2δ (3P arithmetic), so a sub-Gaussian mean test separates them to TV ≥ 1 − ε at finite Φ_d-budget. Deficiency: randomization contracts the $\|\cdot\|_1$ distance, so $\|Q_0 - Q_1\|_1 \le \|P_0^m - P_1^m\|_1 + 2\delta(\mathcal{E}\_q^m, \mathcal{E}\_d(\varepsilon))$; with $\|Q_0 - Q_1\|_1 \ge 2(1-\varepsilon)$ and $\|P_0^m - P_1^m\|_1 \le 2(1 - (1 - \nu_{\Phi_q}(S_t))^m)$, $\delta(\mathcal{E}\_q^m, \mathcal{E}\_d(\varepsilon)) \ge (1 - \nu_{\Phi_q}(S_t))^m - \varepsilon$. (ii) On A_m every nonnegative-weight plug-in $\hat{Z}_m(\Phi_d) = \Sigma_i\, w_i\, e^{\Phi_d S_i}$ (or its self-normalized form) has all support points in S_t^c, so its reconstruction of the band term $\int$_{S_t} e^{Φ_d s} ν(ds) = ν{Φ_d}(S_t)·Z(Φ_d) is identically zero: the fraction ν_{Φ_d}(S_t) of the target is omitted, and the total undershoots by at least that fraction unless the complement term is overestimated. P(A_m) ≥ $(1 - \nu_{\Phi_q}(S_t))^m$ as in (i). No a.s. or in-expectation domination is asserted: unbiased nonnegative-weight estimators exist, so any pathwise-domination claim would be self-contradictory; the content is event-conditional missing mass, with the median form observed in §6.3(iv). ∎

Proposition 3N. Layer-cake, from the 3P′ proof and with its log factor (which cannot be dropped): $1 + \chi^2 \le$ LR_min + sup_{t>0}[ν_{Φ_d}(S_t)²/ν_{Φ_q}(S_t)]·log(LR_max/LR_min), where the LR level sets are top bands and log(LR_max/LR_min) = (Φ_d − Φ_q)·osc(u). LR_min ≤ 1: r = dν_{Φ_d}/dν_{Φ_q} is increasing with E_{ν_{Φ_q}}[r] = 1, so inf r ≤ 1 (strictly < 1 except in the degenerate case; no claim that the same-set sup is ≥ 1 is made or needed: the full-space band attains 1 but violates feasibility (i) and is not in $\mathcal{T}$). Rearranging: sup_{t>0} ν_{Φ_d}(S_t)²/ν_{Φ_q}(S_t) ≥ $(1 + \chi^2 - \text{LR\_min})/((\Phi_d - \Phi_q)\cdot\text{osc}(u))$. Hypothesis (BD) supplies sup_{t∈$\mathcal{T}$_{2δ}} ν_{Φ_d}(S_{t/2})²/ν_{Φ_q}(S_t) ≥ c_geo^{-1}·sup_{t>0} ν_{Φ_d}(S_t)²/ν_{Φ_q}(S_t) — it is doing real work, bridging the feasible-set restriction ($\mathcal{T}$_{2δ} versus all bands) and the half-band numerator (S_{t/2} versus S_t), neither of which follows from the standing assumptions. Task reduction and constant: a test certifying G_w(Φ_d) ≤ δ versus ≥ 2δ with both errors ≤ 1/3 must separate Theorem 3P's pair instantiated at 2δ — $w_0 \equiv 0$ ($G_0 = 0 \le \delta$) versus $w_1 = -h_t$ with h = 8δ/ν_{Φ_d}(S_{t/2}) ($G_1 \ge 2\delta$), admissible exactly for t ∈ $\mathcal{T}$_{2δ}, i.e. conditions (ii)–(iii) read with 8δ. Theorem 3P at 2δ gives n ≥ $(1/36)\sigma^2(2\delta)^{-2}$·sup_{t∈$\mathcal{T}$_{2δ}} ν_{Φ_d}(S_{t/2})²/ν_{Φ_q}(S_t) = $(1/144)\sigma^2\delta^{-2}$·sup_{t∈$\mathcal{T}$_{2δ}} ν_{Φ_d}(S_{t/2})²/ν_{Φ_q}(S_t); inserting (BD): n ≥ $(144\cdot\text{c\_geo})^{-1}\cdot\sigma^2\delta^{-2}\cdot(1 + \chi^2$ − LR_min)/((Φ_d − Φ_q)·osc(u)). At the measured §6.3 geometry (the band_dominance.json calculation, a separate object from the deep-tilt overlap run and unaffected by it): the top-atom (all-band) ratio sup_{t>0} ν_{Φ_d}(S_t)²/ν_{Φ_q}(S_t) = 0.7806²/0.1542 = 3.95, attained in the t → 0 atom limit — this is the right-hand side of (BD), capped by Cauchy–Schwarz at the contemporaneous base-only overlap read $1 + \hat{\chi}^2 = 4.20$ (that base-only overlap read is not the circulation-facing overlap claim; the band-geometry ratios below are unaffected by the deep-tilt run); the feasible-band functional at the calibrated (δ = 0.08, $\hat{L} = 1.336$, $\hat{B} = 3.280$) is 2.90 at t ≈ 0.58, the same value at the 2δ-read and δ-read band sets (condition (ii) excludes t < 0.31, removing the top-atom band — the gap to the second support point is 0.22). So (BD) holds there with c_geo = 3.95/2.90 ≈ 1.36, not near 1; the feasible functional equals the all-band 3.95 only for δ below ≈ 0.029 (2δ-read; ≈ 0.057 at the δ-read). Both

functionals are persisted in the experiment repo (`results/band_dominance.json`). Why the residual cannot be moved to the upper bound by this route: for $\nu \propto e^{\{-2\Phi_d \cdot s\}}$ on [0,1] with $\Phi_q = 0$, $\Phi_d = 6$, one computes $1 + \chi^2 = 3.01$ while sup over all bands $t \in (0, \mathrm{osc}(u)]$ (proper bands; every feasible $\mathcal{T}$ lies there, since condition (i) forces $\nu(S_t) \le 1/4$) of Theorem 3P's half-band-numerator functional $\nu_{\Phi_d}(S_{t/2})^2/\nu_{\Phi_q}(S_t)$ is 0.068 (the S_t-numerator layer-cake functional sups at 1.0 here; $c_{geo} = 1.0/0.068 \approx 14.8$) — a 44× shortfall growing linearly in $\Phi_d \cdot \mathrm{osc}(u)$ — so no absolute c_1 exists for the layer-cake route; whether a different technique yields $n \ge c_1\sigma^2\delta^{-2}(1+\chi^2)$ with absolute c_1 is open. Upper bound: Theorem 3P′ stage (ii), with the achievability constant $\tilde{\sigma}^2$. ∎

Proposition 3G. (i) Worlds, ramp, gap arithmetic, and feasibility are Theorem 3P's verbatim: $w_0 \equiv 0$, $w_1 = -h_t$ with $h_t(s) = \min\{h, L(s-(s^*-t))_+\}$, $h = 4\delta/\nu_{\Phi_d}(S_{t/2})$ admissible for $t \in \mathcal{T}$, both worlds realized with W deterministic given u, $W_i(a) = w_i(u(a))$, so $G_0 = 0$ and $G_1(\Phi_d) \ge \delta$. Per-query KL through the ρ-grader: a query at tilt $\Phi \le \Phi_q$ returns (S, V, C), C the (OC)-measurable side stream. Chain rule on the conditional KL given the past: the S-draw is world-independent (the worlds share (μ_0, u)); C given (past, S) is world-independent (its kernel is fixed across welfare functions and the trajectory laws coincide — Appendix B); V given (past, S = s, C) has law $N(\rho \cdot w_i(s), (1-\rho^2)\sigma_0^2)$ — in the deterministic realization V depends on the queried action only through s, and ξ is independent of everything, so conditioning on C changes nothing. The only nonzero term is the Gaussian verdict KL: $E_{\nu_\Phi}[(\rho \cdot h_t(s))^2]/(2(1-\rho^2)\sigma_0^2) \le \rho^2 h^2 \cdot \nu_{\Phi_q}(S_t)/(2(1-\rho^2)\sigma_0^2)$, using $h_t \le h \cdot 1_{S_t}$ and the monotone tail fact applied conditionally on the past — adaptive tilt choices are covered by the divergence decomposition $KL(P_0^n \| P_1^n) = \Sigma_i\, E_0[KL(P_0(\cdot|\mathcal{F}_{i-1}) \| P_1(\cdot|\mathcal{F}_{i-1}))]$, exactly as in Theorem 3P, with the free (OC) stream contributing zero. Le Cam + Pinsker (both errors $\le 1/3 \Longrightarrow KL \ge 2/9$): $n \cdot \rho^2 h^2 \nu_{\Phi_q}(S_t)/(2(1-\rho^2)\sigma_0^2) \ge 2/9$, and substituting $h = 4\delta/\nu_{\Phi_d}(S_{t/2})$ gives $n \ge (4/9)\cdot(1-\rho^2)\sigma_0^2 \cdot \nu_{\Phi_d}(S_{t/2})^2/(16 \cdot \rho^2 \cdot \delta^2 \cdot \nu_{\Phi_q}(S_t))$ $= (1/36)\cdot((1-\rho^2)/\rho^2)\cdot\sigma_0^2\cdot\delta^{-2}\cdot\nu_{\Phi_d}(S_{t/2})^2/\nu_{\Phi_q}(S_t)$; take the sup over $\mathcal{T}$. Class extensions: post-processing of the verdict only lowers each conditional KL (data processing), and a Gaussian-conditional verdict with welfare-independent (trajectory-measurable) variance $v(s) \ge (1-\rho^2)\sigma_0^2$ and mean separation $\le \rho \cdot h_t(s)$ satisfies the same display by the Gaussian KL formula — with the variance common to the two worlds, the per-hit KL is $(\Delta\mathrm{mean})^2/(2v(s)) \le \rho^2 h_t(s)^2/(2(1-\rho^2)\sigma_0^2)$. Welfare independence of the variance is required, not decorative: a variance depending on W is itself a verdict side-band whose KL the mean-shift formula does not control — between worlds with variances $v_0(s)$, $v_1(s)$ the per-hit KL gains $\tfrac{1}{2}[\log(v_1/v_0) + v_0/v_1 - 1]$, unbounded under a variance floor alone — which would break the converse exactly as non-Gaussian noise would; non-Gaussian noise with merely sub-Gaussian tails is excluded from the class because its shift KL can exceed any Gaussian bound (bounded-support noise under a mean shift has infinite KL). Both exclusions are stated in the Definition. ρ-free floor: off S_t the two worlds' verdict laws coincide at every ρ ($h_t \equiv 0$ there, and the variance is world-independent), so Corollary 3CC's proof runs verbatim with Y replaced by V and the deficiency floor $(1 - \nu_{\Phi_q}(S_t))^m$ holds for all $\rho \in [0,1]$, including the noiseless $\rho = 1$. (ii) ρ known: $V/\rho = W(a) + \sigma_{eff}(\rho)\cdot\xi$ with $\sigma_{eff}(\rho) = \sqrt{(1-\rho^2)\cdot\sigma_0/\rho}$ — this is §5's query channel at noise $\sigma_{eff}(\rho)$, with the estimand w(s), the conditional fluctuation Var(W|u=s), and the range bound B untouched by the rescaling; Theorem 3P′ runs unchanged with $\tilde{\sigma}_{eff}^2 = \sigma_{eff}^2 + \sup_s \mathrm{Var}(W|u=s) + B^2$, and the matching statements (3P′'s attained $O((\Phi_d - \Phi_q)\cdot\mathrm{osc}(u))$ factor and band-localization residual; Proposition 3N under (BD)) carry over with $(\sigma^2, \tilde{\sigma}^2) \to (\sigma_{eff}^2, \tilde{\sigma}_{eff}^2)$, both sides now carrying the same σ_{eff}. $\rho \ge \rho_{min}$ only: rescaling by ρ_{min} gives $V/\rho_{min} = \tilde{W}(a) + \sigma' \cdot \xi$ with $\tilde{W} = (\rho/\rho_{min})\cdot W$ and $\sigma' = \sqrt{(1-\rho^2)\cdot\sigma_0/\rho_{min}} \le \sigma_{eff}(\rho_{min})$ (since $\rho \ge \rho_{min}$); $\tilde{W}$ has range $\le B/\rho_{min}$, conditional variance $\le \rho_{min}^{-2}\cdot\mathrm{Var}(W|u=s)$, and Lipschitz parameter $\le L/\rho_{min}$, so Theorem 3P′ at the ρ_{min}-inflated parameters certifies the rescaled gap $\tilde{G}_w(\Phi_d) = (\rho/\rho_{min})\cdot G_w(\Phi_d)$ — the tilt path of $\tilde{W}$ is pointwise $(\rho/\rho_{min})\cdot T_w$, so the gap scales by the same positive factor. Since $\rho/\rho_{min} \ge 1$, a certificate $\tilde{G} \le \delta$ implies $G_w \le \delta$ (sound); a safe world with $\rho > \rho_{min}$ may fail the certificate (conservative, one-sided). ∎

Dichotomy computations. Polynomial edge $\nu(ds) \sim c_\nu(s^*-s)^\wedge\beta\, ds$ near s^*: with $a = \beta + 1$ and $C = c_\nu\Gamma(a)$, Watson's lemma gives $Z(\Phi) \sim C\, e^{\{\Phi s^*\}\Phi}\{-a\}$. Therefore, at fixed Φ_q, $\chi^2+1 = Z(2\Phi_d - \Phi_q)Z(\Phi_q)/Z(\Phi_d)^2$ $\sim 2^{\{-a\}e}\{-\Phi_q s^*\}Z(\Phi_q)C^{\{-1\}\Phi_d}a$. The constants do not cancel down to the large-cap shortcut unless Φ_q is also sent to infinity; that shortcut is not the fixed-cap display. Gaussian $Z(\Phi) = e^{\{\mu\Phi + s_\nu^2\Phi^2/2\}}$: $\chi^2+1 = e^{\{s_\nu^2(\Phi_d-\Phi_q)^2\}}$ exactly. Essential edge $\nu \asymp e^{\{-c/(s^*-s)\}}$: saddle-point gives $\log Z(\Phi) = \Phi s^* - 2\sqrt{(c\Phi)} + o(\sqrt{\Phi})$, $\chi^2+1 = e^{\{\Theta(\sqrt{\Phi_d})\}}$. ∎

Appendix D: Numerical specification

All quadrature: trapezoid on uniform grids, $\mu \in [0,20]$ with $4\times10^5+1$ points, $\theta \in [0,1]$ with $2\times10^5+1$ points; exponent shifts max-subtracted for stability; environment pinned (numpy). 1-D instance: $E_\Phi[W]$ declines 22.00 $\to$ 10.02 ($\Phi = 500$ proxy for the limit); $\mathrm{Cov}_{\mu_0}(u,W) = -3.606$. 2-D instance ($c = 20$, $\kappa = 60$): $E_0 = 52.00$; peak 74.24 at $\Phi^* = 0.733$; limit 70.01; $\Delta = 4.24$; $\mathrm{Cov}_{\mu_0}(u,W) = +96.39$ (G2 ✓); existence conditions $c\kappa/12 = 100 > 3.606$ and $\kappa/c = 3 < 7.5$ ✓. Certification price in the 1-D geometry ($\Phi_q = 2.2$): $\log(1+\chi^2) = 0.416, 0.857, 1.485$ at $\Phi_d = 5, 10, 30$ — the polynomial branch (cross-validated against an independent implementation to three decimals).

§6.3 experiment specification. Public replication repository: https://github.com/darfaz/goodhart-gap-experiment. The original base bundle used gpt2-medium `6dcaa7a9…`, proxy OpenAssistant/reward-model-deberta-v3-base `7dab4156…`, and gold OpenAssistant/reward-model-deberta-v3-large-v2 `c355404e…`, with `results/samples.csv` containing 6,400 rows (50 prompts × 128 completions, all hashes unique). That base bundle remains the source for the branch-(a) curve, the original coupling result, and the m-sweep. The circulation-facing overlap/one-sidedness claim uses the definitive deep-tilt run reported in Figure 1 and §6.3(iv): $n = 300{,}000$ temperature-tilted samples, $\Phi_q = 2$, $\Phi_d = 3..15$, with diagnostics showing flat but degenerate overlap, ESS near 1, and PSIS-k reliability only through $\Phi_d = 4$. Conventions: u standardized within prompt with population SD; W raw gold logits; $\hat{Z}(\Phi) = m^{-1}\Sigma \exp(\Phi u_i)$ on pooled standardized u; price factor $\hat{Z}(2\Phi_d-\Phi_q)\hat{Z}(\Phi_q)/\hat{Z}(\Phi_d)^2$ ($= 1 + \chi^2(\nu_{\Phi_d}\|\nu_{\Phi_q})$ for empirical tilts; Gaussian closed-form check $\log(1+\chi^2) = \sigma^2(\Phi_d-\Phi_q)^2$ passes). Verification (deriver ≠ referee): base-bundle artifacts regenerate from `results/samples.csv`; the deep-tilt headline diagnostics are summarized in Figure 1 and Figure 2, with a public replication update pending. Gold means are reported in raw DeBERTa-large logit units (within-prompt standardization of W would change finite-sample curves; recorded per the normalization discipline).

Selected references

Agapiou, O., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. M. (2017), "Importance sampling: Intrinsic dimension and computational cost," Statistical Science 32(3), 405–431. DOI:10.1214/17-STS611.

Agrawal, K., Teo, V., Vazquez, J. J., Kunnavakkam, S., Srikanth, V., and Liu, A. (2025), "Evaluating LLM Agent Collusion in Double Auctions," arXiv:2507.01413, ICML 2025 Workshop on Multi-Agent Systems in the Era of Foundation Models.

Armstrong, S. (2018), "Counterfactual equivalence for POMDPs, and underlying deterministic environments," arXiv:1801.03737.

Armstrong, S., and Mindermann, S. (2018), "Occam's razor is insufficient to infer the preferences of irrational agents," Advances in Neural Information Processing Systems 31. arXiv:1712.05812.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1987), Regular Variation, Cambridge University Press.

Brown-Cohen, J., Irving, G., and Piliouras, G. (2024), "Scalable AI safety via doubly-efficient debate," arXiv:2311.14125.

Chatterjee, S., and Diaconis, P. (2018), "The sample size required in importance sampling," Annals of Applied Probability 28(2), 1099–1135. DOI:10.1214/17-AAP1326.

Christiano, P., Cotra, A., and Xu, M. (2021), "Eliciting Latent Knowledge," Alignment Research Center report.

de Bruijn, N. G. (1981), Asymptotic Methods in Analysis, Dover edition.

Le Cam, L. (1986), Asymptotic Methods in Statistical Decision Theory, Springer.

Tsybakov, A. B. (2009), Introduction to Nonparametric Estimation, Springer.

Gao, L., Schulman, J., and Hilton, J. (2023), "Scaling laws for reward model overoptimization," Proceedings of the 40th International Conference on Machine Learning (ICML 2023). arXiv:2210.10760.

Gödel, K. (1931), "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I," Monatshefte für Mathematik und Physik 38, 173–198.

Holmström, B. (1982), "Moral hazard in teams," Bell Journal of Economics 13(2), 324–340. DOI:10.2307/3003457.

Karlin, S. (1968), Total Positivity, Volume I, Stanford University Press.

Karwowski, J., Hayman, O., Bai, X., Kiendlhofer, K., Griffin, C., and Skalse, J. (2024), "Goodhart's Law in Reinforcement Learning," International Conference on Learning Representations (ICLR 2024). arXiv:2310.09144.

Khalaf, H., Verdun, C., Oesterling, A., Lakkaraju, H., and Calmon, F. (2025), "Inference-Time Reward Hacking in Large Language Models," Advances in Neural Information Processing Systems 38 (Spotlight). arXiv:2506.19248.

Khan, S., Saveski, M., and Ugander, J. (2024), "Off-policy evaluation beyond overlap: Sharp partial identification under smoothness," Proceedings of the 41st International Conference on Machine Learning (ICML 2024), PMLR 235. arXiv:2305.11812.

Korbak, T., Perez, E., and Buckley, C. L. (2022), "RL with KL penalties is better viewed as Bayesian inference," Findings of EMNLP 2022. arXiv:2205.11275.

Lang, L., Foote, D., Russell, S., Dragan, A., Jenner, E., and Emmons, S. (2024), "When Your AIs Deceive You: Challenges of Partial Observability in RLHF," Advances in Neural Information Processing Systems 37. arXiv:2402.17747.

Manheim, D., and Garrabrant, S. (2018), "Categorizing Variants of Goodhart's Law," arXiv:1803.04585.

Manski, C. F. (2003), Partial Identification of Probability Distributions, Springer Series in Statistics.

Ng, A. Y., and Russell, S. (2000), "Algorithms for inverse reinforcement learning," Proceedings of the 17th International Conference on Machine Learning (ICML 2000), 663–670.

Tarski, A. (1936), "Der Wahrheitsbegriff in den formalisierten Sprachen," Studia Philosophica 1, 261–405.

Waudby-Smith, I., and Ramdas, A. (2024), "Estimating means of bounded random variables by betting," Journal of the Royal Statistical Society Series B 86(1), 1–27. DOI:10.1093/jrsssb/qkad009.