CALIBRATION CANNOT MOVE THE COST–QUALITY FRONTIER OF AN LLM CASCADE

Anonymous authors
Paper under double-blind review

ABSTRACT

Cascaded inference routes each query to the cheapest model that can answer it, escalating when a confidence score falls below a threshold. A large body of work treats miscalibrated confidence as the bottleneck for that decision and adds a recalibration step. We show this step cannot help. A strictly increasing recalibration map only reparametrizes the score, so it cannot change which queries a threshold accepts at a matched escalation rate, and the achievable cost–quality set is unchanged. This identity holds for arbitrarily many tiers and on every sample, not just on average (Theorem 1). We place it in a four-way taxonomy indexed by the σ-algebra the escalation rule can see. The frontier turns out to depend only on the base rate and the ROC curve; no calibration quantity appears in it (Theorem 2). We measure the invariance on 6,471 generations from three model tiers over 701 items of GSM8K, MATH-500, and MMLU-Pro. Temperature and Platt scaling move the swept frontier by exactly zero, in all 15 task $\times$ signal cells, while mean expected calibration error (ECE) falls by roughly half. The zero survives a $25\times$ sweep of the opus:haiku price ratio. Discrimination is what moves the frontier instead, and it is close to saturated: six confidence signals are statistically indistinguishable, with AUROC between 0.761 and 0.767. A fused score improves AUROC significantly ($+0.044$, 95% CI $[+0.007, +0.083]$) but yields no resolvable frontier gain, a pre-registered prediction we falsify. Calibration's real benefit is narrower than the literature assumes, and one-sided. It lowers how often a label-free risk promise is broken, from 35.8% of cells raw to 13.5% after temperature scaling, without making the risk estimate itself any more accurate, and hitting a target escalation budget needs no calibration at all. Recalibration is a controllability device, not a capability device.

1 INTRODUCTION

A cascade answers each query with a cheap model first, and only calls in an expensive model when the cheap one seems unsure. That decision, whether to escalate, is normally made by thresholding a confidence score. Most work on cascades focuses on making that score numerically accurate: fit a calibrator, such as Platt scaling, isotonic regression, or histogram binning, so that a reported confidence of 0.9 really does mean 90% correct, and then threshold on that corrected scale (Chen et al., 2023; Aggarwal et al., 2023; Kotte, 2026). The implicit assumption is that a badly calibrated score is what makes escalation decisions bad.

That assumption is false, and not just approximately. Recalibration changes units, not order. Escalation only asks whether a score falls above or below a cutoff. Stretching or shifting the scale, as long as it never reverses two scores' order, cannot move a query from one side of that cutoff to the other. So at any fixed escalation rate, the calibrated policy and the raw policy accept exactly the same queries, on the same data, at the same cost and accuracy. Halving the calibration error changes nothing about the cascade's cost–quality frontier, not just on average but query by query, and this holds for cascades with any number of tiers (Theorem 1). This is easy to miss because calibration and escalation look like they should be coupled: a better-calibrated score is a better estimate of how likely the answer is to be correct, and escalation is supposed to be a decision based on exactly that likelihood. But a threshold rule never looks at the value of that estimate, only its rank among all the other scores, and rank is exactly what calibration leaves untouched. What calibration changes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

is something else entirely, and useful for a different reason: it lets a practitioner pick a threshold without needing any labeled data at all.

Several recent cascade papers fit a calibrator before thresholding and credit their reported savings to the pipeline as a whole, calibration step included (Chen et al., 2023; Aggarwal et al., 2023; Kotte, 2026); one reports a cost reduction of over 30% attributed in part to a calibrated threshold rule. Theorem 1 says this attribution cannot be right. A calibration step that only reorders the score cannot be the reason a threshold accepts fewer, or more expensive, queries, because thresholds cannot see the reordering. Whatever is producing these savings, it has to be the underlying signal, the ranking it induces, or the rule used to pick the threshold, not the recalibration step. We make this precise in what follows: a theorem, a direct measurement of it, and an account of what is actually driving the gains the literature has been crediting to calibration.

Contributions. **(i)** We prove an exact invariance theorem: for cascades with any number of tiers, and with no assumptions on the data, a strictly increasing recalibration leaves the achievable cost–quality set exactly unchanged. This holds through an explicit bijection between policies that preserves the accept-or-escalate decision on every single query (Theorem 1). We organize this and related results into a taxonomy based on how much information the escalation rule can see (Table 1), and show why the invariance holds in general: the frontier depends only on the base rate and the ROC curve, and no calibration-specific quantity appears anywhere in it (Theorem 2). **(ii)** We measure the invariance directly. Across $6{,}471$ generations from three model tiers, temperature and Platt scaling move the swept frontier by exactly $0.00\text{e}{+00}$ in all 15 task $\times$ signal cells and all 40 frontier cells we test, while mean ECE falls from 0.0883 to 0.0478 and 0.0394 respectively (§6.1). **(iii)** We show what does move the frontier, including a result we did not expect. Discrimination moves it, and one cheap signal already comes close to saturating it: six of our seven confidence signals are statistically indistinguishable from each other, with AUROC between 0.761 and 0.767. Fusing them out-of-fold raises AUROC significantly ($+0.044$, 95% CI $[+0.007, +0.083]$) but produces no resolvable gain on the frontier itself, a pre-registered prediction we falsify (§6.2). **(iv)** We locate exactly where calibration does help, and state it one-sidedly rather than overselling it. Temperature scaling cuts the rate at which realized risk exceeds its promised bound from 35.75% to 13.52% (-22.24 points, 95% CI $[-23.69, -20.76]$), while the average gap between realized and target risk gets slightly worse. Hitting a target escalation budget, by contrast, needs no calibration at all. And across workloads, the per-item risk guarantee transfers, but the resulting accept rate does not (§6.3–§6.4).

We are not arguing that calibration is useless, nor that any existing cascade system fails to work or that its reported numbers are wrong; §2 returns to the attribution question directly. Nor are we claiming that no transformation can help: per-group recalibration and fusion both can, precisely because they bring in information the original score did not have (Theorem 13).

2 Related work

Cascaded inference for language models began as a cost-reduction architecture with FrugalGPT (Chen et al., 2023), formalized probabilistically by Dohan et al. (2022); Varshney & Baral (2022) study the accuracy/efficiency trade-off directly. AutoMix (Aggarwal et al., 2023) adds self-verification and a meta-router, Hybrid LLM (Ding et al., 2024) learns a router trained on quality gaps, and RouterBench (Hu et al., 2024) provides an evaluation suite; more recent work optimizes the threshold structure itself (Zellinger & Thomson, 2025; Valkanas et al., 2025; Zellinger et al., 2025; Warren & Dras, 2025). Our results are orthogonal to all of this: these papers choose which signal and which policy structure to use, and Theorem 2 says those are exactly the choices that can move the frontier. What none of them can do is move it by improving the score's numerical calibration.

UCCI (Kotte, 2026) and Conformal Cascade (Dou et al., 2026) are the two papers closest to ours, and both place calibration at the center of the escalation decision. UCCI maps token-level margin uncertainty to a per-query error probability with isotonic regression, then picks the escalation threshold by cost minimization, reporting a 31% cost reduction (95% CI $[27\%, 35\%]$) on a production named-entity-recognition workload. Conformal Cascade uses conformal prediction set size as the deferral rule, giving a distribution-free accuracy guarantee across tiers. We do not think either result is wrong. Our claim is narrower: Theorem 1 says a reported cost reduction of this form cannot be *caused* by the calibration step itself, since a strictly increasing recalibration cannot change which queries a

threshold accepts; it must come from the signal, the ranking it induces, or the threshold-selection rule. Conformal Cascade's prediction-set deferral is a coarsening of the score, row (b) of Table 1, so our theory predicts it is frontier-neutral at best relative to thresholding the score directly, while it buys a distribution-free guarantee a raw threshold lacks (Angelopoulos & Bates, 2021). Appendix I gives the full positioning against both papers.

Verbalized confidence (Tian et al., 2023; Xiong et al., 2023) and self-consistency or semantic-entropy signals (Farquhar et al., 2024; Nikitin et al., 2024) are the two practical signal families in use; models have some access to their own correctness (Kadavath et al., 2022). Temperature scaling and its relatives (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005; Zadrozny & Elkan, 2002) remain the standard post-hoc calibration toolkit, and the reliability/resolution decomposition of proper scoring rules (Murphy, 1973; DeGroot & Fienberg, 1983) is the classical statement that these are separate coordinates; our contribution is to show the cascade frontier reads only the resolution coordinate. A cascade tier's accept/reject structure is the classical reject option (Chow, 1970; Geifman & El-Yaniv, 2017), and Lemma 4 is a Neyman-Pearson argument of that kind, though rejection here carries a price rather than an abstention cost. To our knowledge the exact invariance statement in Theorem 1 has not been stated for cascades before.

3 Problem setup

Cascade and policy class. A cascade has tiers $k = 1, \dots, K$ ordered cheap to expensive with per-query cost $c_k > 0$; a query answered at tier k pays the cumulative $C_k \coloneqq \sum_{j \le k} c_j$, since every tier invoked is billed. Queries are i.i.d. from a workload $\mathbb{P}$. Tier k returns a correctness indicator $Y_k \in \{0, 1\}$ and, for $k < K$, a score $S_k \in [0, 1]$; write $\eta_k(s) \coloneqq \mathbb{E}[Y_k \mid S_k = s]$. A *threshold policy* $(\tau_k, \gamma_k)_{k<K}$ accepts at tier k with probability

$$a_k(s) \;=\; \mathbf{1}\{s > \tau_k\} \;+\; \gamma_k \, \mathbf{1}\{s = \tau_k\}, \tag{1}$$

otherwise escalating; tier K always answers. Write $\Phi(S_1, \dots, S_{K-1})$ for this class. The boundary randomization γ_k is not cosmetic: it makes coverage range over all of $[0, 1]$ when the score has atoms, and makes the class closed under the reparametrizations we study — and discrete signals (self-consistency over k samples takes at most $k + 1$ values) make atoms the normal case here.

Frontier. With $\kappa(\pi)$ the random tier that answers, set $\mathcal{C}(\pi) \coloneqq \mathbb{E}[C_{\kappa(\pi)}]$ and $\mathcal{Q}(\pi) \coloneqq \mathbb{E}[Y_{\kappa(\pi)}]$; the *achievable set* is $\mathcal{A}(S_1, \dots, S_{K-1}) \coloneqq \{(\mathcal{C}(\pi), \mathcal{Q}(\pi)) : \pi \in \Phi\}$ and the *frontier* its northwest boundary. Sweeping τ traces $\mathcal{A}$, which is exactly the procedure of §6, and because cost is affine in the escalation rate (Lemma 3) "at matched cost" and "at matched escalation rate" are interchangeable; we compare at matched rate, since the budget is what a deployment constrains.

Recalibrators, and the two coordinates we measure. We study the four in standard use: temperature scaling $s \mapsto \sigma(\sigma^{-1}(s)/T)$ and Platt scaling $s \mapsto \sigma(A\sigma^{-1}(s) + B)$ (Niculescu-Mizil & Caruana, 2005), both strictly increasing when $T > 0$, $A > 0$; isotonic regression (Zadrozny & Elkan, 2002), weakly increasing and creating ties; and histogram binning, which need not be monotone at all since empirical bin means can invert under noise. That four-way split is the taxonomy of §4, and it predicts which calibrators leave the frontier exactly fixed and which fix it only in expectation. We measure calibration by ECE, the binned estimator of $\mathrm{CE}_1 \coloneqq \mathbb{E}|\eta(S) - c(S)|$ for reported confidence c, and discrimination by AUROC together with the ROC curve itself, since Theorem 2 shows pointwise ROC dominance — not AUROC ordering — is what transfers. These are separate coordinates of a proper scoring rule's decomposition into reliability and resolution (Murphy, 1973; DeGroot & Fienberg, 1983); the frontier reads only the second.

Cost model, and a harness artifact we must disclose. Costs are per-query dollars from published per-token prices (Haiku 4.5 \$1/\$5, Sonnet 4.5 \$3/\$15, Opus 4.5 \$5/\$25 per million input/output tokens, read from the vendor's public pricing page on 2026-08-18). The API-reported *input*-token counts for our runs include a fixed platform system prompt of roughly 3.5k tokens that is an artifact of our elicitation harness, not of the cascade under study. Billing it would be doubly wrong: it is not a cost a deployment of this cascade would pay, and because its size and cache state vary by tier it would have *inverted the true price ordering of the tiers*. We therefore price input at the true prompt length, identical across tiers at 143 tokens, and output at measured length, giving \$0.002072

Table 1: The taxonomy proved in this section, indexed by the σ-algebra the escalation rule can see. S is the raw score, U a second signal, Z an observed group label. "Frontier effect" is at the population level at matched cost; measured values are the maximum absolute accuracy deviation from the raw sweep over the 15 task × signal cells (Table 2).

Transformation	Example	Information	Frontier effect (measured)
(a) strictly increasing $g(s)$	temperature, Platt	$\sigma(g(S)) = \sigma(S)$	*exactly* invariant, Thm 1: 0.00e+00
(b) weakly increasing $g(s)$	isotonic, with ties	$\sigma(g(S)) \subseteq \sigma(S)$	coarsens; no gain in expectation, Prop 7: 1.83e−02
(c) any non-monotone $h(s)$	histogram binning	$\sigma(h(S)) \subseteq \sigma(S)$	no gain; ties and noise, Prop 8: 2.28e−02
(d) new information	fusion u; per-group z	$\sigma(S, U) \supsetneq \sigma(S)$	*can* strictly gain, Thm 13; not resolved at this n

per haiku sample and \$0.009299 per opus sample — an opus:haiku ratio of $4.49\times$ (arithmetic in Appendix P). Since that ratio depends on output lengths as well as price, §6.4 sweeps it from $2\times$ to $50\times$ rather than resting on it, and we report cost conclusions as ratios. Absolute dollar figures are consequently lower than a naive reading of the API's reported token counts.

4 Theory: what recalibration can and cannot do

A cascade frontier is a functional of the *information* available to the escalation rule — formally, of a σ-algebra — not of the numerical scale on which that information is reported. Recalibration changes the scale, not the σ-algebra, which yields the taxonomy of Table 1. Proofs, formal assumptions and deferred statements are in Appendices A–C; every counterexample and inequality was verified numerically, in exact rational arithmetic where the claim is exact.

Assumptions. Four hold throughout: **(A1)** queries are i.i.d. and $\mathcal{C}, \mathcal{Q}$ are population expectations, so all frontier claims are population claims (the sole finite-sample claim is Corollary 16); **(A2)** each c_k is constant, independent of which queries reach tier k; **(A3)** the tier-k decision sees the query only through S_k, so side information must be declared as part of the score system — the distinction Theorem 13 exploits; **(A4)** escalation is a selection, not a treatment: $\mathbb{E}[Y_{k+1} \mid S_k = s, \text{escalate}] = \mathbb{E}[Y_{k+1} \mid S_k = s] =: a_{k+1}(s)$. For two tiers write $S := S_1$ and define the *escalation value* $\Delta(s) := \eta_1(s) - a_2(s) = \mathbb{E}[Y_1 - Y_2 \mid S = s]$. No result needs a sign for Δ; some need **MON**, that Δ be nondecreasing — the formal content of "the score is oriented correctly". We flag MON at every use; Theorem 1 does *not* need it.

Two structural lemmas. Cost is a relabelling of coverage: with $\alpha := \mathbb{E}[a_1(S)]$, Lemma 3 gives $\mathcal{C} = c_1 + c_2(1 - \alpha)$ and $\mathcal{Q} = \mathbb{E}[a_1(S)\Delta(S)] + \mathbb{E}[a_2(S)]$, so maximizing quality at fixed cost is maximizing $\mathbb{E}[a_1(S)\Delta(S)]$ at fixed coverage. For a σ-algebra $\mathcal{G}$ let the *information ceiling* $Q^\star(\mathcal{G}; \alpha)$ be the best quality achievable by $\mathcal{G}$-measurable randomized accept rules at coverage α; a Neyman–Pearson argument (Lemma 4) identifies it with the upper α-tail of $\Delta_\mathcal{G} := \mathbb{E}[\Delta \mid \mathcal{G}]$ and shows $\alpha \mapsto Q^\star$ is concave, so the swept frontier is already concave and mixing operating points adds nothing. Hence information monotonicity (Lemma 6): $\mathcal{G} \subseteq \mathcal{H}$ implies $Q^\star(\mathcal{G}; \alpha) \le Q^\star(\mathcal{H}; \alpha)$, with equality for *all* α iff $\Delta_\mathcal{H} = \Delta_\mathcal{G}$ a.s. — the spine of the section, since recalibration produces a function of S, which lives inside $\sigma(S)$ and so can never raise the ceiling.

Theorem 1 (monotone invariance)**.** *Let $K \ge 2$ and $g_k : [0,1] \to [0,1]$ be strictly increasing for each $k < K$ (no continuity assumed). Under (A1)–(A3), $\mathcal{A}(g_1(S_1), \ldots, g_{K-1}(S_{K-1})) = \mathcal{A}(S_1, \ldots, S_{K-1})$. The correspondence is stronger than equality of sets: the bijection $\Psi(\tau, \gamma) = (g(\tau), \gamma)$ maps $\Phi(S)$ onto $\Phi(g(S))$ with corresponding policies having* identical acceptance functions on every query *— hence identical expected cost and quality, identical per-tier escalation rates, and identical realizations on any finite sample under a shared seed.*

The proof is three lines plus one boundary case. Cost and quality depend on the policy only through the acceptance functions, since by (A2)–(A4) the probability a query is answered at tier k

is $\mathbb{E}[a_k(S_k)\prod_{j<k}(1-a_j(S_j))]$, so one tier suffices. Strict monotonicity gives $g(s) > g(\tau) \iff s > \tau$ and $g(s) = g(\tau) \iff s = \tau$; substituting into (1) shows the acceptance function of $\Psi(\tau, \gamma)$ on the transformed score equals, *for every* s, that of (τ, γ) on the raw score. Surjectivity needs one extra step: a threshold $\tau' \notin g([0,1])$ has empty tie set, so its accept region is an up-set realized by $\gamma \in \{0,1\}$ at $\inf\{s : g(s) > \tau'\}$. Consequently temperature scaling, Platt scaling with $A > 0$, any strictly increasing link, strictly increasing isotonic regression, and compositions thereof leave the achievable set unchanged at every cost level (Corollary 5). Two caveats we flag rather than hide. The boundary randomization γ is load-bearing — with weak thresholds only, surjectivity can fail at an atom of S where g jumps, and our self-consistency signals take as few as four distinct values. And the corollary concerns maps on the *scalar* score, excluding Platt scaling with fitted $A < 0$ (strictly decreasing; we report the fitted slope for every fit — all 40 frontier cells have $A > 0$, min 0.313) and multiclass temperature scaling on a logit vector, which can reorder queries with equal max-probability and so belongs to row (d).

Rows (b) and (c). Isotonic regression and histogram binning map distinct scores to a common value, so boundary ties must be broken by something other than the score and the achievable set is undefined until a tie-breaking rule is fixed. Proposition 7 shows $Q_{g(S)}(\alpha) \le Q_S(\alpha)$ for every α and nondecreasing g, with any label-independent tie-breaker's realized quality confined to a band whose expectation is again $\le Q_S(\alpha)$: coarsening buys no gain, only variance. Histogram binning is not even weakly monotone, which might suggest the map can help. It cannot, for a different reason — what matters is measurability, not monotonicity: under MON, $Q_{h(S)}(\alpha) \le Q_S(\alpha)$ for *any* measurable h (Proposition 8). Dropping MON opens the one loophole (Proposition 9), and even there what helps is re-ranking, not rescaling (Appendix E). So "non-monotone, therefore it can help" is a non-sequitur, and the *sign* of an observed effect does not separate information from noise — a noise-driven movement can be positive, and Corollary 16 puts that noise at 0.051 in coverage and 0.112 in accuracy at $m = 701$, an order of magnitude above the deviations we measure.

Theorem 2 (rank sufficiency and ROC dominance)**.** *Assume (A1)–(A4), $K = 2$, constant tier-2 accuracy $a_2(s) \equiv a_2$. Let $p := \mathbb{P}(Y_1 = 1)$ and R be the ROC curve of S for Y_1.* ***(i)*** *$\mathcal{A}(S)$ is a functional of (p, R) alone — equivalently of the law of $(F_S(S), Y_1)$: at false-positive rate x, $\alpha(x) = pR(x) + (1-p)x$, $\mathcal{Q}(x) = pR(x) + (1-\alpha(x))a_2$ and $\mathcal{C}(x) = c_1 + c_2(1-\alpha(x))$. No feature of the score's numerical scale — in particular no calibration quantity — appears.* ***(ii)*** *If two scores for the same tier pair have the same p and ROC curves $R \ge \widetilde{R}$ pointwise, then $Q_S(\alpha) \ge Q_{\widetilde{S}}(\alpha)$ at every matched coverage. Pointwise dominance is what transfers;* AUROC *ordering alone is necessary but not sufficient, since ROC curves can cross with ordered areas.*

Theorem 2 is the sharpest form of the thesis. A proper scoring rule decomposes into reliability (calibration) and resolution (discrimination) terms (Murphy, 1973; DeGroot & Fienberg, 1983), and the cascade frontier is a functional of resolution only: a strictly increasing g leaves resolution untouched while it can drive reliability to zero — which is what a recalibrator is for, and why our calibrators halve mean ECE at exactly zero frontier movement.

Where calibration is load-bearing. These theorems locate calibration's role rather than dismissing it: calibration does not select the frontier, it selects a *parametrization* of it — which is what is needed to hit a target without labels. With c the reported confidence, $\mathcal{R}(\tau) := \mathbb{E}[1 - Y_1 \mid c(S) \ge \tau]$ the true accepted risk, $\widehat{\mathcal{R}}(\tau) := \mathbb{E}[1 - c(S) \mid c(S) \ge \tau]$ its label-free surrogate and $\pi_\tau := \mathbb{P}(c(S) \ge \tau)$, Theorem 11 bounds $|\mathcal{R} - \widehat{\mathcal{R}}| \le \mathrm{CE}_1/\pi_\tau$, so the label-free choice $\tau^\star := \inf\{\tau : \widehat{\mathcal{R}}(\tau) \le r\}$ realizes $\mathcal{R}(\tau^\star) \le r + \mathrm{CE}_1/\pi_{\tau^\star}$, and under exact calibration the per-item guarantee is $\tau = 1 - r$. The $1/\pi_\tau$ factor is not optional, and it predicts a fact we measure: the bound goes vacuous where the accept set collapses (§6.4). Part (b) shows why calibration is the right device and nothing else is — $\tau \mapsto \mathcal{R}(\tau)$ is *not identified* from the unlabeled score distribution, since a strictly increasing g leaves the frontier fixed while relabelling which τ carries which risk. Transfer splits the same way (Proposition 12): the *per-item* guarantee transfers exactly, the accept rate and *aggregate* accepted risk do not. And what *can* move the frontier is an enlargement of the σ-algebra: with U auxiliary and $W := \mathbb{E}[Y_1 - Y_2 \mid S, U]$, Theorem 13 gives $Q_{S,U}(\alpha) \ge Q_S(\alpha)$ with equality for *all* α iff $W = \Delta(S)$ a.s., so the prescription is a conditional-independence test, not a calibration metric. Appendix F expands these points and the general-K scope, which we state but do not claim for the ceiling results.

5 EXPERIMENTAL SETUP

Tasks, tiers, signals. Three short-answer reasoning benchmarks scored by exact match: GSM8K (Cobbe et al., 2021), MATH-500 (Hendrycks et al., 2021; Lightman et al., 2023) and MMLU-Pro (Wang et al., 2024). Three tiers from one model family: Claude Haiku 4.5 ($k = 5$ samples), Sonnet 4.5 ($k = 3$) and Opus 4.5 ($k = 1$), all at temperature 1.0. Three-tier coverage on every item gives 6,471 generations over 719 items; the **analysis set is the** 701 **items** with complete coverage and a matching item record (GSM8K 249, MATH-500 219, MMLU-Pro 233), with tier accuracies 0.840 (haiku first sample), 0.856 (haiku majority of 5), 0.870 (sonnet majority of 3) and 0.902 (opus). Two signal families give seven signals: verbalized confidence elicited in-prompt (Tian et al., 2023; Xiong et al., 2023), summarized as mean, minimum, first-sample and standard deviation; and self-consistency over samples — modal-answer share, negated entropy, negated distinct-answer count (Farquhar et al., 2024; Nikitin et al., 2024) — signs oriented so higher means more confident. Within families the signals are near-identical (Spearman $\rho = 0.9998$ between modal-answer share and negated entropy), so there are effectively *two* signals, not seven, which is why §6.2 reports them as a group. Full protocol in Appendix P.

Protocol. In the targeting experiments every calibrator is fitted on a training split and evaluated on a disjoint split, each replicate drawing an item-level bootstrap resample *and* a fresh stratified 50/50 split, so intervals cover both finite-sample and split variability and comparisons against raw are *paired* within replicate. Fusion is logistic and strictly out-of-fold (5-fold, seed 0). In the frontier tables the Platt map is fitted on the rows it is applied to; we state this rather than fix it, because the quantity at issue is an exact algebraic identity — no split can change a 0.00e+00 difference — and we report the fitted slope for every fit rather than assuming its sign. All intervals are 95% percentile bootstrap over items (2,000 resamples for AUROC, 300 for targeting), differences are paired, and where a difference is smaller than its interval we say so.

Refusals and truncations. These are real features of the data and we disclose rather than silently filter them. Scope matters, so we give both: across all 6,471 generation records **88 came back refusal-flagged and 99 were truncated**; within the 6,309 records of the 701-item analysis set the counts are 64 **and** 68. Refusals fall entirely on the middle tier (64 sonnet, 0 haiku, 0 opus) and concentrate in 53 items (7.6%); truncations are spread across tiers (57 haiku, 7 sonnet, 4 opus) over 33 items. Refusals matter because only 3.1% of refused samples are scored correct against 86.0% of non-refused ones, and sample 0 is the refused one on 25 items, making naive single-sample sonnet accuracy 0.4906 on refusal items against 0.8627 elsewhere. Since the middle tier is used only in the three-tier analysis, we report *four* conventions there rather than choosing one silently: refusals kept, resample-on-refusal, 3-sample majority vote, and refusal items excluded ($n = 648$). Resampling is the deployable fix (25 items need one retry, none exhausts all three, accuracy rises 3.3 pp for a 2.6% cost increase) and is our primary convention. Every two-tier result uses haiku and opus only and is untouched by sonnet refusals.

6 RESULTS

6.1 R1: THE INVARIANCE, MEASURED

We sweep the escalation threshold on all 15 task × signal cells, before and after each calibrator, and record the largest accuracy deviation from the raw sweep (Table 2; per-cell values in Table 5). For the two strictly monotone calibrators the answer is 0.00e+00 — every cell, every escalation rate. This is not a small effect reported as zero; it is the double-precision zero Theorem 1 predicts, because the calibrated policy escalates the identical set of queries. Over the same cells mean ECE falls from 0.0883 to 0.0478 (temperature) and 0.0394 (Platt), improving in 12/15 and 11/15 cells: calibration error moves by a factor of two, frontier position by exactly nothing. Table 3 re-verifies this independently as accuracy-at-matched-rate rather than swept curves — maximum |calibrated − raw| accuracy over 40 frontier cells is 0.00e+00, every fitted Platt slope positive (min 0.313), so Corollary 5 applies as stated rather than via the decreasing-map loophole.

The coarsening calibrators do deviate — up to 1.83×10^{-2} (isotonic) and 2.28×10^{-2} (histogram) — as Proposition 7 predicts, and the movement is boundary tie-breaking rather than information:

Table 2: **R1: the invariance, measured.** Aggregated over all 15 task × signal cells (raw mean ECE 0.0883); per-cell values in Table 5. "Δ accuracy" is the deviation of swept frontier accuracy from the raw sweep. The strictly monotone calibrators deviate by *exactly* zero in every cell while cutting ECE; the coarsening calibrators deviate by amounts Corollary 16 places well inside the $m = 701$ noise band and §6.1 attributes to boundary tie-breaking. †Histogram ECE is zero to machine precision *by construction* (in-sample fit; each bin reports its own observed frequency), not evidence that it is the best calibrator. Source: `invariance.csv`.

Calibrator	mean ECE	ECE improved	max |Δ accuracy|	mean |Δ accuracy|
temperature scaling	0.0478	12/15	0.00e+00	0.00e+00
Platt scaling	0.0394	11/15	0.00e+00	0.00e+00
isotonic regression	0.0496	15/15	1.83e−02	7.59e−03
histogram binning	$< 10^{-15\dagger}$	15/15	2.28e−02	1.36e−02

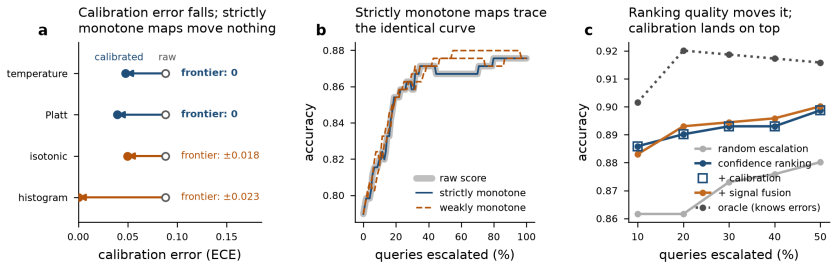


Figure 1: **Recalibrating a confidence score moves its calibration error but not the cascade's cost–quality frontier. (a)** Mean ECE (15 equal-width bins) before and after each map, averaged over the 15 task × signal cells; the annotation is the largest absolute change in swept-frontier accuracy the same map produces in any cell. **(b)** One representative cell's swept frontier: the strictly monotone curve lies exactly on the raw curve at every sweep point and is drawn thinner so the raw curve stays visible beneath it. The weakly monotone curve departs only where the map creates ties among items with different labels — boundary tie-breaking, not added information, as the envelope check in §6.1 establishes. **(c)** Accuracy at fixed escalation rates for five policies: ranking quality moves the frontier (random, to confidence ranking, to the error-knowing oracle), while adding calibration on top of confidence ranking lands on the confidence-ranking curve. Escalation is to the top tier and accuracy is the mean over all 701 items in every policy, so the curves are directly comparable; the item is the unit of replication throughout.

against the best/worst tie-breaking envelope for every (signal, calibrator, rate) cell, **raw accuracy lies inside the envelope in 69 of 70 cells and calibrated accuracy in 70 of 70** (Appendix J), an order of magnitude inside the $m = 701$ noise band.

6.2 R2–R4: DISCRIMINATION MOVES THE FRONTIER, AND FUSION IS A NULL

AUROC for predicting tier-1 correctness, pooled, runs from 0.767 [0.718, 0.816] (modal-answer share) down to 0.761 [0.705, 0.814] (*first-sample* verbalized confidence) across six signals, with the standard deviation of verbalized confidence trailing at 0.670 [0.607, 0.727] (Table 4). Two things matter, and the second is usually mishandled. First, every signal is informative — all seven bootstrap lower bounds exceed 0.5 — and ranking on a single raw signal moves the frontier substantially, which is what Theorem 2 requires. Second, **the ordering among the top six is not real and we do not present it as one**: their spread is 0.006 while the narrowest single interval is 0.097 wide, and the two self-consistency variants have Spearman $\rho = 0.9998$ with each other. These are two families of indistinguishable quality, not a leaderboard, and the cheapest signal — one sample — is inside the pack.

Fusion improves ranking and does not move the frontier. Theorem 13 says fusion *can* move the frontier, and we pre-registered the expectation that the fused policy would sit strictly between the best single raw signal and the oracle at every rate. **It does not, and we report the null.**

The ranking improvement is real: out-of-fold logistic fusion reaches AUROC 0.819 (verbalized + self-consistency) and 0.824 (all six) against 0.767 for the best single signal, a paired-bootstrap difference of $+0.044$, 95% CI $[+0.007, +0.083]$, one-sided $p = 0.009$. The frontier improvement is not: on the pooled $k = 5$ rows of Table 3 the differences at $10/20/30/40/50\%$ escalation are $-0.29/ + 0.29/ + 0.14/ + 0.29/ + 0.14$ pp (averaged over all eight cells at each rate: $-0.25/ + 0.34/ + 0.35/ + 0.40/ + 0.35$ pp), *every* bootstrap interval straddles zero (widest $[-0.86, +1.50]$ pp), the mean over all 40 frontier cells is $+0.237$ pp, fusion beats raw in 22 of 40, and at 10% escalation it is *below* raw. Fusion's frontier gain is at most a few tenths of a point and unresolvable at $n = 701$ — which is not the same as fusion not working.

The damping has an arithmetic explanation, and it is the most transferable number we report: escalation can only help where the strong tier repairs the weak one, and $\mathbb{P}(\text{opus correct} \mid \text{haiku majority wrong}) = 0.4554$ ($n = 101$) against $\mathbb{P}(\text{opus wrong} \mid \text{haiku majority right}) = 0.0233$ ($n = 600$), so an escalation slot is worth at most 0.455 accuracy points, not 1 — and fusion's $+0.071$ precision gain at 20% escalation predicts only $+0.68$ pp, below the noise floor at this n (Appendix K). **A large frontier gain from better confidence scoring on a task pair with a sub-50% repair rate is arithmetically forbidden**, so cascade papers should report this rate. Ranking closes 41–61% of the random-to-oracle gap across those rates and fused out-of-fold 47–56% (per-rate values in Appendix).

6.3 R5–R6: WHAT CALIBRATION BUYS IS GUARANTEE COMPLIANCE, NOT TRACKING OR BUDGET

If calibration cannot move the frontier, what is it for? Theorem 11 says: choosing a threshold that meets a risk target *without labels*. Measured on haiku verbalized confidence, pooled, over attainable binding targets $r \in \{0.05, 0.10, 0.15\}$ with paired bootstraps against raw, the benefit is real but one-sided in a way the folklore is not. A practitioner who treats an uncalibrated score as a probability and asks for risk $\leq r$ gets a threshold that breaks its promise on 35.8% of cells; after temperature scaling it breaks it on 13.5% ($\Delta = -0.2224$, $[-0.2369, -0.2076]$), Platt 16.3%, isotonic 20.9%, histogram 22.2% — every calibrator significantly better, and the effect replicates on two other tier/signal pairs (Appendix L). But the usually-claimed benefit — realized risk landing *near* the target — does not appear: mean $|\text{realized} - \text{target}|$ gets significantly *worse* for every calibrator (raw 0.0384 against temperature 0.0451, $\Delta = +0.0066$ $[+0.0058, +0.0074]$), because the conservative shift that suppresses violations also pushes realized risk below target, and over all 24 cells at attainable binding targets **no calibrator improves absolute tracking in aggregate and the ranking among them is unstable** (Appendix L). We expected calibration to improve tracking; it does not — the second of our two falsified predictions. Calibration is a conservatism device, not a targeting device.

Budget targeting, by contrast, needs no calibration at all: a target escalation rate is a quantile of the score, and a quantile is rank-based, so any strictly increasing map gives the identical escalation set. Across all 168 cells mean absolute escalation-rate error is 0.17587 raw against 0.17588 for temperature scaling — identical to four decimals — while the coarsening calibrators are worse (0.19437 isotonic, 0.23717 histogram) and deviate from raw's realized rate by up to 0.289 against 0.0013 for temperature. Both residual nonzero deviations for strictly monotone maps are diagnosed in Appendix O: **calibration buys risk control, not budget control**.

6.4 R7–R9: TRANSFER, PRICE, AND THE COST OF THE GATE

Three further results, with full tables in Appendices N–G. *Transfer* shows the asymmetry Proposition 12 predicts: applying a τ fitted on task A to task B, the mean accept-rate shift (0.104–0.174) is $3\times$ the mean risk mismatch (0.049–0.086), so **a practitioner sizing their strong-tier budget from the calibration workload's accept rate can be off by a factor of two or more**. *Price* separates two conflated claims: over a $2\times$–$50\times$ sweep the largest frontier-AUC deviation from any strictly monotone calibrator is 1.2×10^{-3} — not the exact zero of Tables 2–3, for a reason we have not isolated, but *numerically identical at all ten price ratios* and below the $1/701$ single-item frontier resolution — while the cost-optimal escalation rate moves from 0.489 at $2\times$ to 0.000 at $30\times$. *The gate can cost more than the answer*: five haiku samples cost \$0.00979 per query, 105% of answering everything with opus (\$0.00930), so the $k = 1$ regime (\$0.00207, 22.3% of all-opus) is the only economically live one — and first-sample verbalized confidence matches the five-sample signals in AUROC. The middle tier does not earn its place either: three-tier is indistinguishable from or worse

than two-tier above 0.88 accuracy under every refusal convention, and in the 0.87–0.88 band its cost advantage has an undetermined sign (95% CI $[-49.0\%, +39.4\%]$).

Structural limits on what these experiments can show. Three limits are properties of the data, not the methods, and reporting numbers past them would be reporting artifacts (details in Appendix H). At $\geq 10\%$ acceptance the minimum risk an oracle threshold can reach is 0.031 for haiku verbalized confidence pooled, so $r = 0.02$ **is unattainable** pooled for every signal and calibrator, the accept set is *empty* in 27–53% of replicates there, and the apparent "excellent" realized risks there **must not be quoted as successes** — as Theorem 11(a) predicts. Conversely any r above the tier's marginal error rate (0.144 haiku pooled) is met by accepting everything, so pooled $r \geq 0.15$ is **non-binding**; we flag `binding` and exclude those rows from every headline number. And with $n = 219$–249 per task no per-task risk difference below 0.03 is resolvable.

7 Limitations

Sample size, and the fusion null. The analysis set is 701 items, 219–249 per task: enough to make the invariance claim decisively — it is an algebraic identity, and $0.00\text{e}{+00}$ does not become nonzero with more data — but not enough to resolve small frontier effects, since the $m = 701$ band is 0.051 in coverage and 0.112 in accuracy while the fusion effect we seek is a few tenths of a point. Our fusion result should therefore be read as *unresolved at this* n, not as evidence of absence.

One model family, one signal type, one task type. All three tiers come from a single vendor's family, and all our signals are verbalized confidence and self-consistency over samples: we have *no logit-based signals*. Our absolute AUROC figures may therefore be pessimistic relative to systems with logit access, and the one case where the invariance is an open empirical question rather than a measured zero — multiclass temperature scaling on a full logit vector, row (d) of Table 1 — is the case we could not test. The tasks are short-answer benchmarks with exact-match grading; nothing here tests open-ended generation, multi-turn interaction or agentic tool use. The theory needs only a binary correctness indicator and a scalar score, but the repair rate (0.455) and the price ladder are properties of this workload.

Costs, pooling, and in-sample fitting. Our dollar figures rest on the cost model of §3, including the deliberate exclusion of the harness's platform system prompt — the right call, but it makes absolute figures lower than a naive reading of the API's token counts and the cascade-versus-opus-only comparison depends on it; the published prices were read on a single date (see the unverified-claims register). Pooled numbers are not a deployment: scoring each item by its task's base accuracy alone already reaches pooled AUROC 0.657, so pooled figures borrow leverage from between-task base-rate differences. Finally, the Platt map in Table 3 is fitted in-sample and the best raw signal is chosen by AUROC on the rows the frontier is measured on: the first cannot affect an exact algebraic identity, and the second mildly favours raw over fused, making our fusion null *conservative*. The general-K ceiling results are stated but not proved.

8 Conclusion

A cascade's escalation rule is a threshold, and a threshold reads rank, not value; recalibration changes value, not rank. That is why the invariance is exact rather than approximate: strictly monotone recalibration leaves the achievable cost–quality set pointwise identical, for any number of tiers and without distributional assumptions, and we measured it at $0.00\text{e}{+00}$ in every cell we swept while ECE fell by half. Calibration still buys something real, just not that: it lets a practitioner set a threshold from unlabeled data and keep a per-item risk promise across a workload shift, though only for risk, not for budget. What actually moves the frontier is discrimination, and it is capped by how often the expensive tier repairs the cheap tier's mistakes, 0.455 in our data, a number cascade papers should report alongside ROC.

ETHICS STATEMENT

This work involves no human subjects, no personally identifying data, and no new data collection from people. The three benchmarks used (GSM8K, MATH-500, MMLU-Pro) are public research datasets used as published, for their intended evaluation purpose. Our generations come from commercially available model APIs under their terms of service; we release the generation records and analysis code so that the results can be checked without re-incurring API cost.

Two ethical considerations are specific to the content of the paper. First, the paper's constructive claim concerns *risk guarantees*, and we have been careful not to overstate it: a cascade that promises "at most r error on accepted queries" after post-hoc calibration is making a promise that our own measurements show is violated on 13.5% of cells even after the best calibrator we tried, and whose absolute accuracy is no better than the uncalibrated score's. Deploying such a guarantee in a setting where an error is costly — clinical, legal, financial — would require validation we have not performed, and §6.4 documents regimes ($r = 0.02$ pooled) where the guarantee is not merely loose but unattainable, and where an apparently excellent realized risk is an artifact of a near-empty accept set. We flag these rather than reporting the favourable-looking numbers. Second, cascade research is cost-reduction research, and cost reduction lowers the barrier to deploying language models at scale in applications where accuracy matters; our results are neutral on whether any particular such deployment is appropriate, and the accuracy ceilings we report (no cascade point at or below strong-model cost reaches strong-model accuracy) should temper rather than encourage claims that a cascade is a free lunch.

We used a model family from a single vendor and report measured prices from that vendor's public pricing page. We have no relationship to disclose that would affect that choice, and the price-ratio sweep of §6.4 is included precisely so that no conclusion in the paper depends on one vendor's price list.

REPRODUCIBILITY STATEMENT

Every number in this paper is traceable to a released artifact. The theory is self-contained: all assumptions are stated in §4 and Appendix A, Theorem 1 and Theorem 2 are proved in the main text, and the remaining proofs are in Appendix C. Every counterexample and inequality was additionally verified numerically before being written, in exact rational arithmetic where the claim is exact; those checks and their outputs are listed in Appendix D.

For the experiments, the elicitation protocol (tasks, tiers, sample counts, temperature, prompt structure) is in §5 and Appendix P; the cost model, including the harness token-accounting artifact and how we handled it, is in §3; the handling of the 88 refusal-flagged and 99 truncated sonnet samples, and the four conventions we report rather than choosing one, is in §5. The per-cell result tables underlying each results subsection are released as CSVs and named in the caption of every table (`invariance.csv`, `decomposition.csv`, `frontier_table.csv`, `targeting.csv`, `transfer.csv`, `price_sweep.csv`, `three_tier_table.csv`), together with the raw generation records for all $6{,}471$ generations, so that any reported figure can be recomputed without re-running inference. An anonymized repository containing the elicitation harness, the analysis code, the raw records and the table-generating scripts will be provided with the submission; we do not link it here to preserve double-blind review.

Two reproducibility caveats we state rather than hide. The frontier tables fit the Platt map on the rows it is applied to and select the best single signal by AUROC on the rows the frontier is measured on (§7); the first cannot affect an exact algebraic identity and the second makes our fusion null conservative. And the published per-token prices underlying all dollar figures were read on a single date and have not been independently re-verified; they are listed in the unverified-claims register released with the supplementary material, along with every other claim in this paper that was not checked against a primary source.

USE OF LARGE LANGUAGE MODELS

We disclose AI involvement in this work in full, against the required and recommended task categories of the ICLR 2027 AI Policy for Authors. This disclosure is unusually extensive because the AI in-

volvement was unusually extensive: **an AI system performed essentially all of the research work in this paper under human direction**, and describing it as an editing aid would be a misrepresentation.

Tasks with required disclosure — all performed with AI assistance. An AI system: *proposed and refined the central hypothesis* (it proposed the inverted thesis of this paper after finding that the project's original hypothesis was both anticipated by existing work and contradicted by our own pilot measurements; the human authors chose this thesis from the options it presented); *developed the theoretical framework* (the σ-algebra taxonomy of Table 1 is its formulation); *formulated the mathematical claims and wrote the proofs* of every theorem, proposition, lemma and corollary in §4 and Appendix C, including supplying the load-bearing technical ingredients (the monotone-value assumption 4, without which Proposition 8 is false; the boundary randomization in (1), without which Theorem 1 fails at atoms; and the $1/\pi_\tau$ factor in (4), without which Theorem 11(a) is false); *designed the experiments and the methodology*, including the four-calibrator split that mirrors the taxonomy, the paired-bootstrap protocol, and the tie-breaking-envelope diagnostic; *implemented the methods*, comprising the elicitation harness, all calibration and targeting code, and all analysis code; *cleaned and reformatted the data*, including the refusal and truncation handling of §5; and *interpreted the results*, including the repair-rate mechanism of §6.2 and the one-sided reading of §6.3. No synthetic datasets were generated for training or evaluation purposes; the model generations analysed here are the object of study rather than a synthetic dataset, and no AI-assisted translation was involved.

Tasks with recommended disclosure — also performed with AI assistance. The same system created and edited all software code, ran the experiments, created the figures, drafted all parts of this paper including this statement, summarized and analysed the existing literature and identified the relevant prior work, searched for information, proposed the paper's structure and title, formatted the references, and edited the text for readability. It also suggested the experimental parameters (escalation-rate grid, target-risk grid, bootstrap replicate counts, price-ratio sweep range).

Self-correction, and errors it caught. We record this because it is material to how much weight a reader should place on the pipeline. During the work the system identified and corrected three of its own implementation errors — a logit-domain bug in the calibration code, a diverging Platt fit on a low-cardinality signal, and the token-accounting artifact described in §3 which, uncorrected, would have inverted the tiers' price ordering — and it falsified two of its own pre-registered predictions, reporting both as nulls in this paper: that signal fusion would produce a frontier gain strictly between raw and oracle (§6.2), and that calibration would improve absolute risk tracking (§6.3). It also corrected two claims it had itself previously asserted, which are now stated in their corrected and weaker forms with assumptions explicit: taxonomy row (c) is false without the monotone-value assumption (MON, Assumption 4) (Proposition 9 gives the counterexample), and threshold transfer across workloads is false for aggregate quantities and true only for the per-item guarantee (Proposition 12).

Human responsibility and verification. The human authors directed the work, chose the thesis from the options offered, set the standard of evidence, and are responsible for the final content of this paper — including every claim, number, proof and artifact produced with AI assistance. We have reviewed all AI-assisted work. Verification performed: the theory's counterexamples and inequalities were re-derived numerically and independently of the prose (Appendix D); every number in the main text was checked against the released per-cell CSVs; every citation in the bibliography was retrieved from the arXiv or Crossref API rather than recalled, and claims we could not verify against a primary source are listed in the unverified-claims register released with the supplementary material rather than presented as confirmed. Where a result did not replicate or an interval straddled zero, we report the null. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, Shyam Upadhyay, Manaal Faruqui, and Mausam. Automix: Automatically mixing language models. *arXiv preprint arXiv:2310.12963*, 2023. URL `https://arxiv.org/abs/2310.12963`.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. URL https://arxiv.org/abs/2107.07511.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023. URL https://arxiv.org/abs/2305.05176.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32(1–2):12–22, 1983. doi: 10.2307/2987588.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid llm: Cost-efficient and quality-aware query routing. *arXiv preprint arXiv:2404.14618*, 2024. URL https://arxiv.org/abs/2404.14618.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A. Saurous, Jascha Sohl-dickstein, Kevin Murphy, and Charles Sutton. Language model cascades. *arXiv preprint arXiv:2207.10342*, 2022. URL https://arxiv.org/abs/2207.10342.

Yifan Dou, Shikan Lian, and Shibo Li. Conformal cascade: Distribution-free accuracy guarantees for multi-tier llm inference. *arXiv preprint arXiv:2607.25018*, 2026. URL https://arxiv.org/abs/2607.25018.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *arXiv preprint arXiv:1705.08500*, 2017. URL https://arxiv.org/abs/1705.08500.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017. URL https://arxiv.org/abs/1706.04599.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. URL https://arxiv.org/abs/2103.03874.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. Routerbench: A benchmark for multi-llm routing system. *arXiv preprint arXiv:2403.12031*, 2024. URL https://arxiv.org/abs/2403.12031.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Varun Kotte. Ucci: Calibrated uncertainty for cost-optimal llm cascade routing. *arXiv preprint arXiv:2605.18796*, 2026. URL https://arxiv.org/abs/2605.18796.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973. doi: 10.1175/1520-0450(1973)012¡0595:ANVPOT¿2.0.CO;2.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, 2005. doi: 10.1145/1102351.1102430.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *arXiv preprint arXiv:2405.20003*, 2024. URL https://arxiv.org/abs/2405.20003.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.330.

Antonios Valkanas, Soumyasundar Pal, Pavel Rumiantsev, Yingxue Zhang, and Mark Coates. C3po: Optimized large language model cascades with probabilistic cost constraints for reasoning. *arXiv preprint arXiv:2511.07396*, 2025. URL https://arxiv.org/abs/2511.07396.

Neeraj Varshney and Chitta Baral. Model cascading: Towards jointly improving efficiency and accuracy of nlp systems. *arXiv preprint arXiv:2210.05528*, 2022. URL https://arxiv.org/abs/2210.05528.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024. URL https://arxiv.org/abs/2406.01574.

David Warren and Mark Dras. Bi-directional model cascading with proxy confidence. *arXiv preprint arXiv:2504.19391*, 2025. URL https://arxiv.org/abs/2504.19391.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023. URL https://arxiv.org/abs/2306.13063.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002. doi: 10.1145/775047.775151.

Michael J. Zellinger and Matt Thomson. Rational tuning of llm cascades via probabilistic modeling. *arXiv preprint arXiv:2501.09345*, 2025. URL https://arxiv.org/abs/2501.09345.

Michael J. Zellinger, Rex Liu, and Matt Thomson. Cost-saving llm cascades with early abstention. *arXiv preprint arXiv:2502.09054*, 2025. URL https://arxiv.org/abs/2502.09054.

Table 3: **R1/R3/R4: the frontier decomposition.** Pooled ($n = 701$), $k = 5$ elicitation, majority-vote answer retained at tier 1 (accuracy 0.8559, \$0.00979/query); tier 2 is opus $k = 1$ (accuracy 0.9016, \$0.00930/query). Each cell is accuracy / \$ per query at the stated escalation rate. Best single raw signal is `verb_mean`, selected by AUROC; calibrated is Platt on the same signal (fitted slope $0.823 > 0$). **Calibrated equals raw to** $0.00\text{e}{+00}$ **at every rate**, which is Theorem 1 restated on the frontier. Fusion is out-of-fold and does not sit strictly between raw and oracle: it is *below* raw at 10%. Last column: percent of the random-to-oracle gap closed, raw / fused. Source: `frontier_table.csv`.

Escal.	(a) random	(b) best raw	(c) calibrated	(d) fused OOF	(e) oracle	closed (%)
10%	0.8616 / 0.01067	0.8859 / 0.01138	**0.8859** / 0.01138	0.8830 / 0.01128	0.9016 / 0.01107	60.7 / 53.6
20%	0.8616 / 0.01173	0.8902 / 0.01269	**0.8902** / 0.01269	0.8930 / 0.01261	0.9201 / 0.01193	48.8 / 53.7
30%	0.8730 / 0.01269	0.8930 / 0.01382	**0.8930** / 0.01382	0.8944 / 0.01384	0.9187 / 0.01242	43.7 / 46.9
40%	0.8759 / 0.01354	0.8930 / 0.01485	**0.8930** / 0.01485	0.8959 / 0.01482	0.9173 / 0.01291	41.4 / 48.3
50%	0.8802 / 0.01447	0.8987 / 0.01590	**0.8987** / 0.01590	0.9001 / 0.01570	0.9158 / 0.01354	52.0 / 56.0

Table 4: **R2: discrimination, pooled ($n = 701$).** AUROC for predicting tier-1 (haiku majority-of-5) correctness, with 95% percentile bootstrap intervals over items ($2{,}000$ resamples). Every lower bound exceeds 0.5, so no signal is at chance; but the spread across the top six (0.006) is an order of magnitude smaller than any single interval width, so **their ordering is not resolvable and we do not report one** — the table is sorted only for legibility. Source: `decomposition.csv`.

Signal	Family	AUROC	95% CI	CI width
modal-answer share (`agree`)	self-consistency	0.767	$[0.718, 0.816]$	0.098
negated answer entropy (`entropy`)	self-consistency	0.766	$[0.718, 0.816]$	0.098
mean verbalized (`verb_mean`)	verbalized	0.763	$[0.705, 0.815]$	0.110
minimum verbalized (`verb_min`)	verbalized	0.762	$[0.705, 0.814]$	0.109
negated distinct answers (`n_distinct`)	self-consistency	0.762	$[0.713, 0.810]$	0.097
first-sample verbalized (`verb_first`)	verbalized	0.761	$[0.705, 0.814]$	0.109
verbalized std. dev. (`verbstd`)	verbalized	0.670	$[0.607, 0.727]$	0.119

Table 5: **Per-cell version of Table 2.** Number of distinct score values, ECE before and after recalibration, and the maximum absolute deviation of swept-frontier accuracy from the raw sweep, for each of the 15 task × signal cells. Source: `invariance.csv`.

			ECE			max $\lvert\Delta$ accuracy$\rvert$ vs. raw sweep			
Task	Signal	#vals	raw	temp.	Platt	temp.	Platt	isotonic	histogram
GSM8K	verb_mean	39	0.0216	0.0330	0.0223	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$4.02\text{e}{-03}$	$8.03\text{e}{-03}$
	verb_first	10	0.0398	0.0510	0.0135	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$4.02\text{e}{-03}$	$8.03\text{e}{-03}$
	verb_min	12	0.0364	0.0584	0.0248	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$8.03\text{e}{-03}$	$8.03\text{e}{-03}$
	agree	4	0.0352	0.0182	0.0093	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$1.61\text{e}{-02}$
	neg_entropy	4	0.0401	0.0199	0.0112	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$1.61\text{e}{-02}$
MATH-500	verb_mean	53	0.1329	0.0508	0.0380	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$1.37\text{e}{-02}$	$1.83\text{e}{-02}$
	verb_first	11	0.1365	0.0591	0.0420	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$1.37\text{e}{-02}$	$2.28\text{e}{-02}$
	verb_min	12	0.1380	0.0524	0.0305	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$1.83\text{e}{-02}$	$1.37\text{e}{-02}$
	agree	6	0.1262	0.0253	0.0137	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$4.57\text{e}{-03}$	$2.28\text{e}{-02}$
	neg_entropy	9	0.1076	0.0197	0.0095	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$4.57\text{e}{-03}$	$1.83\text{e}{-02}$
MMLU-Pro	verb_mean	99	0.1090	0.0963	0.1212	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$8.58\text{e}{-03}$	$1.29\text{e}{-02}$
	verb_first	17	0.0898	0.0847	0.0937	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$8.58\text{e}{-03}$	$8.58\text{e}{-03}$
	verb_min	21	0.0972	0.0853	0.0984	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$8.58\text{e}{-03}$	$8.58\text{e}{-03}$
	agree	6	0.1015	0.0264	0.0262	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$4.29\text{e}{-03}$	$8.58\text{e}{-03}$
	neg_entropy	9	0.1122	0.0361	0.0366	$0.00\text{e}{+00}$	$0.00\text{e}{+00}$	$1.29\text{e}{-02}$	$1.29\text{e}{-02}$
mean			0.0883	0.0478	0.0394	$0.00\text{e}{+00}$		$7.59\text{e}{-03}$	$1.36\text{e}{-02}$

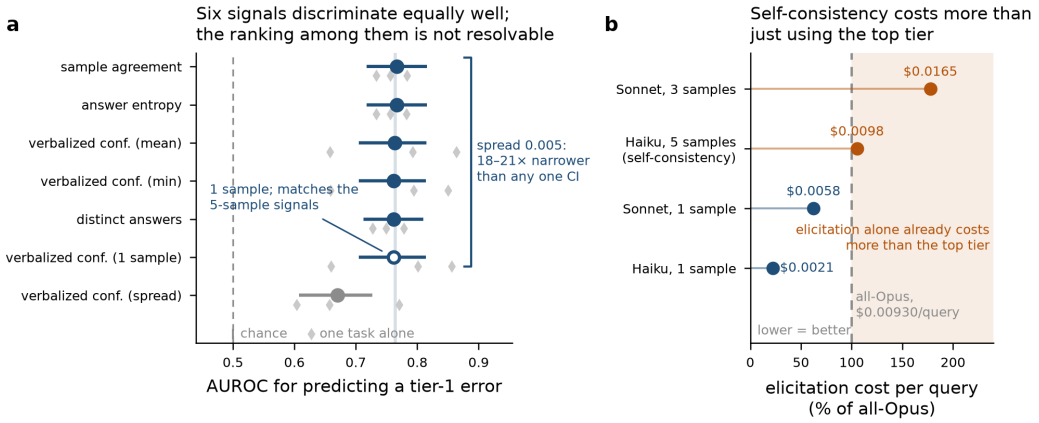


Figure 2: **Six of seven confidence signals discriminate tier-1 errors indistinguishably well, and the five-sample signals cost more to elicit than simply answering with the top tier. (a)** AUROC for predicting whether the bottom tier's first sample is wrong, pooled over all 701 items (base accuracy 0.840), with 95% percentile bootstrap intervals resampled over items; grey diamonds show the same signal within each task separately ($n = 249/219/233$) and are context only, not pooled into the interval. The tinted band spans the six leading signals, AUROC 0.761 (verbalized confidence from a single sample) to 0.767 (sample agreement). That 0.006 spread is roughly a tenth of the width of any single interval (0.097–0.110), so the ordering within the band is not resolvable from these data and we do not interpret it; the bracket marks this explicitly. Every lower bound exceeds 0.5, so no signal is at chance, and the one clearly separated signal — the spread of verbalized confidence across samples, AUROC 0.670 $[0.607, 0.727]$ — is the only ranking statement the data support. The open marker is the single-sample signal, which needs one generation rather than five and matches the self-consistency signals. **(b)** Measured elicitation cost per query as a percentage of answering every query with the top tier ($0.00930/query, dashed line); dollar values are means over the 701 items. Five-sample self-consistency on the bottom tier costs $0.00979/query, 105% of all-Opus, and three-sample Sonnet 178%, so the shaded region contains regimes where the uncertainty signal alone costs more than the cascade it is meant to replace. Only the single-sample bottom-tier regime ($0.00207, 22.3% of all-Opus) leaves room to save money — and by (a) its signal is as discriminative as the expensive ones.

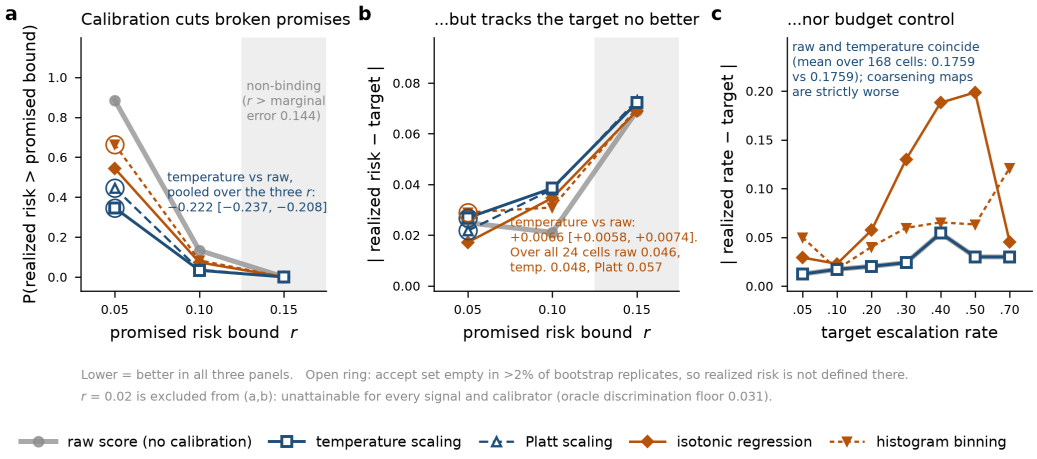


Figure 3: **Calibration buys one-sided compliance with a risk promise; it does not buy better absolute risk tracking, and it does not buy budget control.** Label-free targeting on the bottom tier's verbalized confidence, pooled over all 701 items: a threshold is chosen so the calibrated score's predicted risk on the accepted set meets a target, no labels are used to choose it, and realized risk or realized escalation rate is then measured. Panels share the same five conditions — the raw score and four recalibration maps fitted on the same data — and lower is better in all three. Intervals are 95% percentile bootstrap over items (600 resamples for the plotted cell). **(a)** Fraction of resamples in which realized risk exceeds the promised bound. Pooled over $r \in \{0.05, 0.10, 0.15\}$ the violation rate falls from 0.3575 (raw) to 0.1352 (temperature), a paired difference of -0.2224 $[-0.2369, -0.2076]$; Platt 0.1626, isotonic 0.2090, histogram 0.2221. The effect replicates on Sonnet verbalized confidence (-0.176 $[-0.191, -0.162]$) and Haiku sample agreement (-0.293 $[-0.301, -0.285]$). **(b)** Mean absolute distance between realized risk and target, same conditions. Here calibration is slightly *worse*: temperature minus raw is $+0.0066$ $[+0.0058, +0.0074]$, and across all 24 tier $\times$ signal $\times$ scope cells the means are raw 0.046, temperature 0.048, Platt 0.057, isotonic 0.044, histogram 0.057, with an unstable ranking. Panels (a) and (b) together are the claim: the benefit is one-sided violation control, not accuracy of the risk estimate. **(c)** Mean absolute error in hitting a target *escalation rate*. Raw and temperature scaling coincide to four decimals (0.17587 vs 0.17588 over 168 cells; the curves overlap, which is why raw is drawn thicker), while the coarsening maps are strictly worse (0.19437 isotonic, 0.23717 histogram; maximum deviation from raw's realized rate 0.289 and 0.277 against 0.0013 for temperature). Platt scaling sits between at 0.17723: it is strictly monotone, so Theorem 1 applies to it, and its small excess over raw comes from the two implementation effects diagnosed in Appendix O, not from coarsening. **Exclusions.** $r = 0.02$ is omitted from (a, b): it is unattainable for every signal and calibrator (oracle discrimination floor 0.031 for this cell), the accept set is empty in 27–53% of resamples there, and the resulting "excellent" realized risks are artifacts of near-empty accept sets rather than successes. Grey bands mark $r \geq 0.15$, non-binding pooled because the marginal error rate is 0.144. Open rings mark points where the accept set was empty in more than 2% of resamples, so realized risk is defined on a shrinking subset — at $r = 0.05$ this affects the calibrated conditions and not raw, which is itself part of the mechanism and not a nuisance.

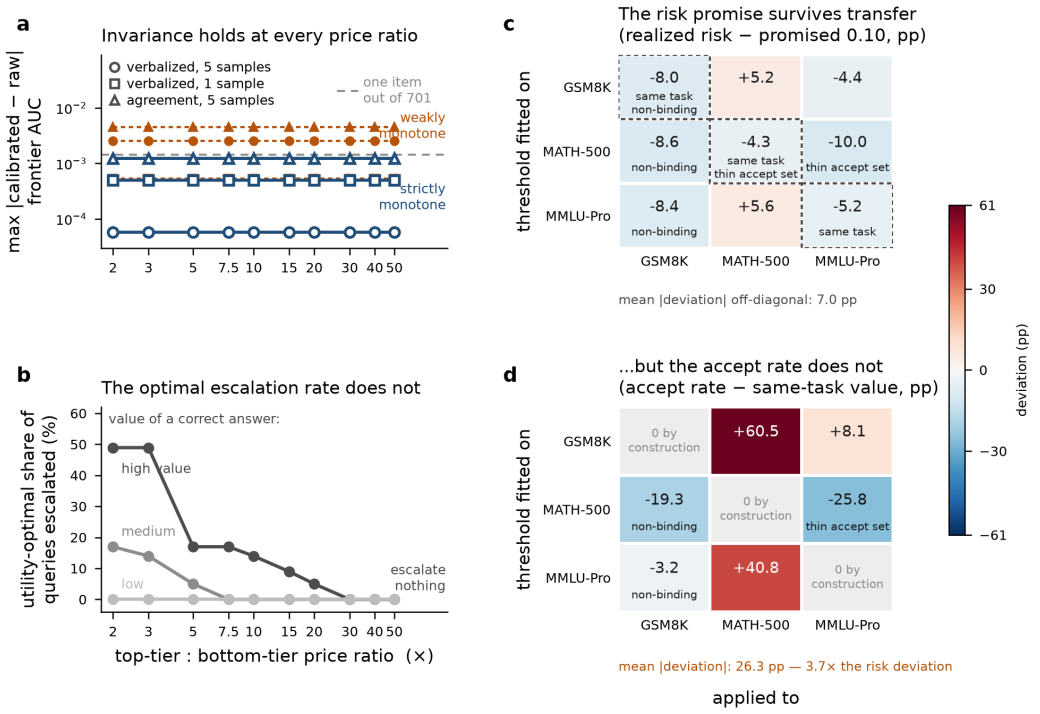


Figure 4: **Invariance is a property of the ranking, not of the prices; the transferred risk promise survives a workload change but the accept rate does not.** **(a)** Largest absolute difference in swept-frontier AUC between a calibrated and the raw escalation policy as the top-tier:bottom-tier price ratio sweeps $2\times$ to $50\times$ (our measured ladder is $4.5\times$). Each marker shape is one signal; open blue markers are the strictly monotone maps and filled amber the weakly monotone ones. Every series is flat: the deviation is a property of the map, not the price. Over the whole sweep the strictly monotone maximum is $5.7\text{e}{-05}$ for five-sample verbalized confidence, $5.0\text{e}{-04}$ for single-sample verbalized confidence and $1.2\text{e}{-03}$ for sample agreement — all below the $1/701 = 1.4\text{e}{-03}$ frontier resolution of a single item (dashed line), i.e. below what one item's reclassification would move. The weakly monotone maxima ($5.3\text{e}{-04}$ to $4.5\text{e}{-03}$) exceed that resolution for two of three signals, consistent with the tie-breaking band of Figure 1(b). **(b)** The same sweep's utility-optimal escalation rate for three values of a correct answer ($V = 0.20, 0.05, 0.01$ per query, $k = 1$), moving from 49% down to 0% as the top tier gets more expensive: the optimum is price-dependent, the invariance in (a) is not. **(c, d)** Cross-workload transfer: a calibrator and threshold are fitted on one task at promised bound $r = 0.10$ and applied unchanged to another, one cell per ordered pair with $n = 249/219/233$ on the evaluated task. **(c)** Realized risk minus the promised 0.10 in percentage points; the promise is roughly kept, mean absolute deviation 7.0 pp over the six off-diagonal cells, bound exceeded in two (fit on GSM8K or MMLU-Pro, evaluated on MATH-500: $+5.2$ and $+5.6$ pp). **(d)** Accept rate minus the accept rate the same target produces in-distribution on the evaluated task; mean absolute deviation 26.3 pp, $3.7\times$ the risk deviation, with cells as large as $+60.5$ pp. Diagonal cells in (c) are the same-task reference and drawn dashed; in (d) they are zero by construction and left blank rather than plotted as a result. Cells are annotated where the promised bound is non-binding for the evaluated task or where the accept set is thin, so those numbers are read as flagged rather than clean. The aggregate over all 390 binding transfer cells with $r \geq 0.05$ shows the same asymmetry: mean |realized risk − target| 5.26 pp against a mean |accept-rate shift| of 12.52 pp for temperature scaling.

A FORMAL STATEMENT OF THE STANDING ASSUMPTIONS

These are the assumptions abbreviated (A1)–(A4) in §4.

Assumption 1 (i.i.d. workload)**.** *Queries are i.i.d. from $\mathbb{P}$ and $\mathcal{C}$, $\mathcal{Q}$ are population expectations. All frontier statements are population statements; the only finite-sample claim is Corollary 16.*

Assumption 2 (decision-independent costs)**.** *Each c_k is a deterministic constant, independent of which queries reach tier k and of the accept/escalate decision. The cost of computing S_k is folded into c_k. In particular a query's cost depends on the policy only through the tier that answers it.*

Assumption 3 (score-measurable policies)**.** *The decision at tier k may depend on the query only through S_k (and k). Any additional side information used by the decision rule must be declared as part of the score system; this is precisely the distinction that Theorem 13 exploits.*

Assumption 4 (monotone value, MON)**.** *$\Delta(s)$ is nondecreasing in s. Used by Propositions 7 and 8 and Theorem 13(i);* not *used by Theorem 1.*

Assumption 5 (no interference)**.** *For each k, $\mathbb{E}[Y_{k+1} \mid S_k = s, \text{escalate}] = \mathbb{E}[Y_{k+1} \mid S_k = s] =: a_{k+1}(s)$: the downstream tier's accuracy on a query does not depend on whether the policy chose to escalate it, only on the query itself. Equivalently, escalation is a selection, not a treatment.*

B STATEMENTS DEFERRED FROM THE MAIN TEXT

Lemma 3 (cost–coverage equivalence)**.** *Under Assumptions 1–5 with $K = 2$, let $\alpha := \mathbb{E}[a_1(S)]$ be the coverage. Then $\mathcal{C} = c_1 + c_2(1-\alpha)$ and $\mathcal{Q} = \mathbb{E}[a_1(S)\Delta(S)] + \mathbb{E}[a_2(S)]$. Consequently $\mathcal{C}$ is a strictly decreasing affine function of α alone, and the (cost, quality) achievable set is the image of the (coverage, quality) achievable set under a fixed bijection; every statement about one transfers verbatim to the other.*

Lemma 4 (Neyman–Pearson ceiling and its attainment)**.** *Let $\mathcal{G}$ be a σ-algebra with Δ integrable and write $\Delta_{\mathcal{G}} := \mathbb{E}[\Delta \mid \mathcal{G}]$. For every $\alpha \in [0,1]$, $Q^\star(\mathcal{G};\alpha) = \int_{1-\alpha}^1 F^{-1}_{\Delta_{\mathcal{G}}}(u)\,\mathrm{d}u + \mathbb{E}[a_2(S)] = \alpha\,\mathrm{CVaR}_\alpha(\Delta_{\mathcal{G}}) + \mathbb{E}[a_2(S)]$, and the supremum is attained by thresholding $\Delta_{\mathcal{G}}$: with $t_\alpha := F^{-1}_{\Delta_{\mathcal{G}}}(1-\alpha)$ and γ chosen so that $\mathbb{E}[a^\star] = \alpha$, the rule $a^\star = \mathbf{1}\{\Delta_{\mathcal{G}} > t_\alpha\} + \gamma \mathbf{1}\{\Delta_{\mathcal{G}} = t_\alpha\}$ is optimal, and any optimal rule agrees with $a^\star$ almost surely on $\{\Delta_{\mathcal{G}} \neq t_\alpha\}$. Moreover $\alpha \mapsto Q^\star(\mathcal{G};\alpha)$ is concave.*

Corollary 5 (standard recalibrators cannot move a cascade frontier)**.** *Under Assumptions 1–3, none of the following changes the achievable set or the frontier of a cascade, at any cost level:* (i) *temperature scaling $s \mapsto \sigma(\sigma^{-1}(s)/T)$ with $T > 0$ and σ the logistic function;* (ii) *Platt scaling $s \mapsto \sigma(A\sigma^{-1}(s) + B)$ with $A > 0$;* (iii) *beta calibration and any parametric link strictly increasing in s;* (iv) *isotonic regression in the strictly increasing case (no pooled blocks);* (v) *any composition of the above.*

Proof. Each map is strictly increasing on $[0,1]$: σ and σ^{-1} are strictly increasing, and $x \mapsto x/T$ ($T > 0$) and $x \mapsto Ax + B$ ($A > 0$) are strictly increasing; compositions of strictly increasing maps are strictly increasing. Apply Theorem 1. □

Lemma 6 (information monotonicity)**.** *If $\mathcal{G} \subseteq \mathcal{H}$ then $Q^\star(\mathcal{G};\alpha) \le Q^\star(\mathcal{H};\alpha)$ for every α, with equality for* all *α if and only if $\Delta_{\mathcal{H}} = \Delta_{\mathcal{G}}$ almost surely.*

Proposition 7 (coarsening and the tie-breaking band)**.** *Assume 1–5,* MON, *$K = 2$. Let g be nondecreasing with level sets (blocks) $B_1, \dots, B_M$, so $\sigma(g(S)) = \sigma(\beta(S))$ for the block index β, and let $Q_S, Q_{g(S)}$ be the best quality at coverage α over threshold policies on S and on $g(S)$.*

(i) ***(No gain.)*** *$Q_{g(S)}(\alpha) \le Q_S(\alpha)$ for every α, with equality for all α iff Δ is almost surely constant within every block.*

(ii) ***(Tie-breaking band.)*** *For any label-independent tie-breaker $V \perp (S, Y_1, Y_2)$, the quality $Q^V_{g(S)}(\alpha)$ realized by a threshold on $(g(S), V)$ satisfies $Q^-(\alpha) \le Q^V_{g(S)}(\alpha) \le Q_S(\alpha)$ with*

$$Q_S(\alpha) - Q^-(\alpha) \;\le\; \mathbb{P}(S \in B_{m(\alpha)}) \cdot \mathrm{osc}_{B_{m(\alpha)}}(\Delta), \tag{2}$$

where $B_{m(\alpha)}$ is the block containing the coverage-α boundary, and $\mathbb{E}_V[Q^V_{g(S)}(\alpha)] = Q_{g(S)}(\alpha) \le Q_S(\alpha)$: in expectation over the tie-break there is no gain.

Proposition 8 (the $\sigma(S)$ ceiling)**.** *Assume 1–5,* MON*, $K = 2$. For* any *measurable $h : [0,1] \to \mathbb{R}$ — monotone, non-monotone, or discontinuous — and every α, $Q_{h(S)}(\alpha) \le Q_S(\alpha)$. The same holds for randomized $h(S,V)$ with $V \perp (S, Y_1, Y_2)$.*

Proposition 9 (when a non-monotone map genuinely helps)**.** *Drop* MON*. Then Proposition 8 can fail strictly: with S uniform on $\{0.1, 0.5, 0.9\}$, tier 2 perfect and $\eta = (0.1, 0.9, 0.5)$, at coverage $\alpha = 1/3$ thresholding S achieves $5/6$ while the reordering that ranks 0.5 above 0.9 achieves $29/30 > 5/6$, which equals the ceiling $Q^\star(\sigma(S); 1/3)$. Any such gain is exactly a gain in* discrimination*: the optimum over $\sigma(S)$ is thresholding $\Delta(S)$, and the improvement is the ROC improvement of $h(S)$ over S in the sense of Theorem 2.*

Theorem 10 (rank sufficiency, restated with the parametrization labelled)**.** *Under the hypotheses of Theorem 2, at false-positive rate x the swept operating point is*

$$\alpha(x) = pR(x) + (1-p)x, \qquad \mathcal{Q}(x) = pR(x) + (1-\alpha(x))a_2, \qquad \mathcal{C}(x) = c_1 + c_2(1-\alpha(x)). \tag{3}$$

Theorem 11 (label-free risk targeting)**.** *Assume 1–5, $K = 2$. Let c be the reported confidence, $\mathcal{R}(\tau) := \mathbb{E}[1 - Y_1 \mid c(S) \ge \tau]$, $\widehat{\mathcal{R}}(\tau) := \mathbb{E}[1 - c(S) \mid c(S) \ge \tau]$ and $\pi_\tau := \mathbb{P}(c(S) \ge \tau)$.* ***(a) Targeting bound:*** *for any τ with $\pi_\tau > 0$,*

$$\left|\mathcal{R}(\tau) - \widehat{\mathcal{R}}(\tau)\right| \;\le\; \frac{\mathbb{E}\left[|\eta(S) - c(S)|\,\mathbf{1}\{c(S) \ge \tau\}\right]}{\pi_\tau} \;\le\; \frac{\mathrm{CE}_1}{\pi_\tau}, \tag{4}$$

so $\tau^\star := \inf\{\tau : \widehat{\mathcal{R}}(\tau) \le r\}$ — computable without labels — realizes $\mathcal{R}(\tau^\star) \le r + \mathrm{CE}_1/\pi_{\tau^\star}$. If $c = \eta$ a.s. then $\mathrm{CE}_1 = 0$, realized risk equals the target, and the per-item guarantee takes the closed form $\tau = 1 - r$. The realized escalation rate $\mathbb{P}(c(S) < \tau)$ is computable without labels for any τ. ***(b) Non-identification without calibration:*** *for strictly increasing g, c and $g \circ c$ induce the same achievable set (Theorem 1) and satisfy $\mathcal{R}_{g\circ c}(g(\tau)) = \mathcal{R}_c(\tau)$, but the risk attached to a fixed numerical threshold changes arbitrarily with g: the map $\tau \mapsto \mathcal{R}(\tau)$ is* not identified *from the unlabeled distribution of the reported score. Calibration is exactly the choice of scale that makes the label-free surrogate $\widehat{\mathcal{R}}$ agree with $\mathcal{R}$.*

Proposition 12 (threshold transfer across workloads)**.** *Let the score be calibrated on both workloads, $\eta_A(s) = \eta_B(s) = c(s)$ a.e.* ***(i)*** *At $\tau = 1 - r$ every accepted query on either workload has conditional error probability $1 - c(S) \le r$; this is workload-free and requires no assumption on the marginals.* ***(ii)*** *The accept rate π_τ^W and the* aggregate *accepted risk do* not *transfer: both are functionals of the law of $c(S)$ under W and can differ arbitrarily even under exact calibration on both, with $|\mathcal{R}^A(\tau) - \mathcal{R}^B(\tau)| \le \mathrm{TV}(\mathcal{L}_A(c(S)), \mathcal{L}_B(c(S)))/\min(\pi_\tau^A, \pi_\tau^B)$.*

Theorem 13 (fusion and conditional recalibration)**.** *Assume 1–5, $K = 2$. Let U be any auxiliary signal (a second confidence score, a group label, a query feature) and define $W := \mathbb{E}[Y_1 - Y_2 \mid S, U]$, so that $\Delta(S) = \mathbb{E}[W \mid S]$. Let $Q_{S,U}(\alpha) := Q^\star(\sigma(S,U); \alpha)$. Then:*

- *(i) $Q_{S,U}(\alpha) \ge Q^\star(\sigma(S); \alpha) \ge Q_S(\alpha)$ for all α, the last inequality being equality under* MON*.*
- *(ii)* ***(Exact condition for strict improvement.)*** *$Q_{S,U}(\alpha) = Q^\star(\sigma(S); \alpha)$ for* all *$\alpha \in [0,1]$ if and only if $W = \Delta(S)$ almost surely, i.e. iff $\mathbb{E}[Y_1 - Y_2 \mid S, U]$ is a function of S alone. When tier-2 accuracy does not depend on (s, u), the condition reduces to $\mathbb{P}(Y_1 = 1 \mid S, U)$ not being $\sigma(S)$-measurable.*
- *(iii)* ***(Attainment.)*** *The improvement is realized by thresholding the fused score W itself, or any strictly increasing function of it, which is the optimal one-dimensional summary of (S, U) in the sense of Lemma 4.*
- *(iv)* ***(Per-group recalibration.)*** *Let Z be an observed group label and g_z a family of strictly increasing per-group maps. The pooled score $g_Z(S)$ is in general* not *a monotone function of S, and thresholding it is a $\sigma(S, Z)$-measurable policy; by (ii) it strictly improves the frontier for some α iff $\mathbb{E}[Y_1 - Y_2 \mid S, Z] \ne \Delta(S)$ with positive probability. If Z is* not *observed at inference time, the policy is not admissible under Assumption 3 and no improvement is available.*

C DEFERRED PROOFS

Theorem 1 is proved in the main text. Proofs below are for Lemmas 3–6, Propositions 7–9, Theorems 2–13 and Proposition 12.

PROOF OF LEMMA 3 (COST–COVERAGE EQUIVALENCE)

Every query pays c_1; a query is additionally charged c_2 exactly when it is escalated, which happens with probability $1-\alpha$, giving the cost identity by linearity (Assumption 2). For quality, condition on S and use Assumption 5: $\mathcal{Q} = \mathbb{E}[a_1(S)\eta(S) + (1-a_1(S))a_2(S)] = \mathbb{E}[a_1(S)(\eta(S)-a_2(S))] + \mathbb{E}[a_2(S)]$, which is the claim by the definition of Δ. Since $c_2 > 0$, $\alpha \mapsto c_1 + c_2(1-\alpha)$ is a strictly decreasing affine bijection. □

PROOF OF LEMMA 4 (NEYMAN–PEARSON CEILING)

For any $\mathcal{G}$-measurable a the tower property gives $\mathbb{E}[a\Delta] = \mathbb{E}[a\Delta_{\mathcal{G}}]$, so we may replace Δ by $\Delta_{\mathcal{G}}$. Let a be $\mathcal{G}$-measurable with $\mathbb{E}[a] = \alpha$ and let $a^\star = \mathbf{1}\{\Delta_{\mathcal{G}} > t_\alpha\} + \gamma \mathbf{1}\{\Delta_{\mathcal{G}} = t_\alpha\}$ with $t_\alpha = F^{-1}_{\Delta_{\mathcal{G}}}(1-\alpha)$ and γ chosen so that $\mathbb{E}[a^\star] = \alpha$. Then

$$\mathbb{E}[a^\star\Delta_{\mathcal{G}}] - \mathbb{E}[a\Delta_{\mathcal{G}}] = \mathbb{E}[(a^\star - a)(\Delta_{\mathcal{G}} - t_\alpha)] + t_\alpha(\mathbb{E}[a^\star] - \mathbb{E}[a]) = \mathbb{E}[(a^\star - a)(\Delta_{\mathcal{G}} - t_\alpha)], \quad (5)$$

the second term vanishing as both rules have coverage α. The integrand is pointwise nonnegative: on $\{\Delta_{\mathcal{G}} > t_\alpha\}$, $a^\star = 1 \geq a$ and the factor is positive; on $\{\Delta_{\mathcal{G}} < t_\alpha\}$, $a^\star = 0 \leq a$ and the factor is negative; on $\{\Delta_{\mathcal{G}} = t_\alpha\}$ it is zero. Hence $a^\star$ is optimal, with equality iff $(a^\star - a)(\Delta_{\mathcal{G}} - t_\alpha) = 0$ a.s., which is the uniqueness claim. A valid γ exists because $t \mapsto \mathbb{P}(\Delta_{\mathcal{G}} > t)$ has left limits, so the atom at t_α absorbs the deficit. The value $\mathbb{E}[a^\star\Delta_{\mathcal{G}}]$ is the integral of the top-α quantiles of $\Delta_{\mathcal{G}}$, which is the stated formula. Concavity: the derivative of $\alpha \mapsto \int_{1-\alpha}^1 F^{-1}(u)\,\mathrm{d}u$ is $F^{-1}(1-\alpha)$, nonincreasing in α. □

PROOF OF LEMMA 6 (INFORMATION MONOTONICITY)

Inclusion of policy classes gives "$\leq$", since every $\mathcal{G}$-measurable rule is $\mathcal{H}$-measurable. For the equality case, $\Delta_{\mathcal{G}} = \mathbb{E}[\Delta_{\mathcal{H}} \mid \mathcal{G}]$ by the tower property, so $\Delta_{\mathcal{G}} \leq_{\mathrm{cx}} \Delta_{\mathcal{H}}$ in the convex order. Suppose the ceilings agree for all α. Then $\int_{1-\alpha}^1 F^{-1}_{\Delta_{\mathcal{G}}} = \int_{1-\alpha}^1 F^{-1}_{\Delta_{\mathcal{H}}}$ for all α; differentiating in α gives equality of the quantile functions Lebesgue-a.e. on $(0,1)$, so $\Delta_{\mathcal{G}}$ and $\Delta_{\mathcal{H}}$ have the same distribution and in particular the same second moment. Since $\mathbb{E}[\Delta_{\mathcal{H}} \mid \mathcal{G}] = \Delta_{\mathcal{G}}$, the conditional variance decomposition $\mathrm{Var}(\Delta_{\mathcal{H}}) = \mathrm{Var}(\Delta_{\mathcal{G}}) + \mathbb{E}[\mathrm{Var}(\Delta_{\mathcal{H}} \mid \mathcal{G})]$ forces $\mathbb{E}[\mathrm{Var}(\Delta_{\mathcal{H}} \mid \mathcal{G})] = 0$, i.e. $\Delta_{\mathcal{H}}$ is $\mathcal{G}$-measurable and equals $\Delta_{\mathcal{G}}$ a.s. The converse is immediate. □

PROOF OF PROPOSITION 7 (COARSENING AND THE BAND)

(i) Every threshold policy on $g(S)$ is a $\sigma(g(S))$-measurable randomized rule and $\sigma(g(S)) \subseteq \sigma(S)$. By Lemma 4 with $\mathcal{G} = \sigma(S)$, the ceiling $Q^\star(\sigma(S);\alpha)$ is attained by thresholding $\Delta(S)$; under MON Δ is nondecreasing in s, so its up-sets are up-sets of S and the optimizer is itself a threshold policy on S of the form (1). Hence $Q_S(\alpha) = Q^\star(\sigma(S);\alpha)$: on the raw score, thresholding is unimprovable within $\sigma(S)$. Then $Q_{g(S)}(\alpha) \leq Q^\star(\sigma(g(S));\alpha) \leq Q^\star(\sigma(S);\alpha) = Q_S(\alpha)$ by Lemma 6. Equality for all α requires $\mathbb{E}[\Delta \mid S] = \mathbb{E}[\Delta \mid \beta(S)]$ a.s., i.e. $\Delta(S)$ is a.s. block-constant.

(ii) Upper bound: a threshold on $(g(S), V)$ is $\sigma(S,V)$-measurable, and $V \perp (S, Y_1, Y_2)$ gives $\mathbb{E}[\Delta \mid S, V] = \Delta(S)$, so by Lemmas 4 and 6 the ceiling over $\sigma(S,V)$ equals that over $\sigma(S)$, namely $Q_S(\alpha)$. Lower bound and width: such a threshold accepts all blocks above $B_{m(\alpha)}$, no block below, and a V-selected sub-population of $B_{m(\alpha)}$ of fixed mass; all such rules agree outside $B_{m(\alpha)}$, so any two differ by at most $\mathbb{P}(S \in B_{m(\alpha)})\,\mathrm{osc}_{B_{m(\alpha)}}(\Delta)$, which is (2). Finally, averaging the realized rule over V yields a rule constant on each block with the same coverage — a randomized threshold policy on $g(S)$ — and by linearity of $a \mapsto \mathbb{E}[a\Delta]$ its expected quality equals that policy's quality, at most $Q_{g(S)}(\alpha)$; conversely $Q_{g(S)}(\alpha)$ is achieved by a uniform tie-break. □

PROOF OF PROPOSITION 8 (THE $\sigma(S)$ CEILING)

$h(S)$ is $\sigma(S)$-measurable, so any threshold policy on $h(S)$ is a $\sigma(S)$-measurable randomized rule. By the argument in Proposition 7(i), MON implies $Q_S(\alpha) = Q^\star(\sigma(S);\alpha)$, the supremum over *all* such rules, giving the claim. For randomized $h(S,V)$, independence gives $\mathbb{E}[\Delta \mid S, V] = \Delta(S)$, so Lemma 6 yields $Q^\star(\sigma(S,V);\alpha) = Q^\star(\sigma(S);\alpha)$. □

PROOF OF PROPOSITION 9 (THE LOOPHOLE)

With $a_2 \equiv 1$, Lemma 3 gives $\mathcal{Q} = \mathbb{E}[a(S)(\eta(S) - 1)] + 1$, so maximizing quality at coverage α means accepting the mass with the largest η. At $\alpha = 1/3$, thresholding S accepts $\{S = 0.9\}$ and yields $\frac{1}{3}(0.5) + \frac{2}{3}(1) = 5/6$; accepting $\{S = 0.5\}$ instead yields $\frac{1}{3}(0.9) + \frac{2}{3}(1) = 29/30$. Since $\Delta = \eta - 1$ is maximized at $s = 0.5$, the latter is the Δ-threshold rule and attains $Q^\star(\sigma(S); 1/3)$ by Lemma 4. For the final clause, the best achievable over $\sigma(S)$ is thresholding $\Delta(S)$, and h improves on S only insofar as the up-sets of $h(S)$ align better with those of $\Delta(S)$; the resulting change in accepted-correct mass at matched coverage is exactly the quantity governed by ROC ordering in Theorem 2. □

PROOF OF THEOREM 2 (RANK SUFFICIENCY AND ROC DOMINANCE)

(i) Let the policy accept the up-set with false-positive rate $x = \mathbb{P}(\text{accept} \mid Y_1 = 0)$ and true-positive rate $R(x) = \mathbb{P}(\text{accept} \mid Y_1 = 1)$; boundary randomization (1) makes x range over $[0, 1]$. Coverage is $\alpha = pR(x) + (1 - p)x$ by total probability and the accepted-correct mass is $pR(x)$. By Lemma 3 with $a_2(s) \equiv a_2$, quality is the accepted-correct mass plus a_2 times the escalated mass, giving (3). Since (p, R) determines $(\mathcal{C}(x), \mathcal{Q}(x))$ for every x, it determines $\mathcal{A}(S)$. The identification of (p, R) with the law of $(F_S(S), Y_1)$ is standard: R is invariant under strictly increasing reparametrization and is a function of the class-conditional distributions of $F_S(S)$.

(ii) Fix α and set $m(x) = pR(x) + (1 - p)x$, $\widetilde{m}(x) = p\widetilde{R}(x) + (1 - p)x$; both are continuous and strictly increasing, hence invertible onto $[0, 1]$. Let $\widetilde{x} = \widetilde{m}^{-1}(\alpha)$, $x = m^{-1}(\alpha)$. Since $R \geq \widetilde{R}$, $m(\widetilde{x}) \geq \widetilde{m}(\widetilde{x}) = \alpha = m(x)$, so $x \leq \widetilde{x}$. Equating coverages,

$$pR(x) + (1 - p)x = \alpha = p\widetilde{R}(\widetilde{x}) + (1 - p)\widetilde{x} \implies pR(x) - p\widetilde{R}(\widetilde{x}) = (1 - p)(\widetilde{x} - x) \geq 0. \quad (6)$$

So the accepted-correct mass of S at coverage α is at least that of $\widetilde{S}$; by (3) the two quality values differ from these masses by the common term $(1-\alpha)a_2$, and costs coincide at matched α by Lemma 3. For the final clause, AUROC is $\int_0^1 R$, and two ROC curves can cross while having ordered areas, in which case (6) applies only on the sub-range where pointwise dominance holds. □

PROOF OF THEOREM 11 (LABEL-FREE RISK TARGETING)

(a) Write $A = \{c(S) \geq \tau\}$. By the tower property $\mathcal{R}(\tau) = \mathbb{E}[(1 - \eta(S))\mathbf{1}_A]/\pi_\tau$ and $\widehat{\mathcal{R}}(\tau) = \mathbb{E}[(1-c(S))\mathbf{1}_A]/\pi_\tau$. Subtracting, $|\mathcal{R}-\widehat{\mathcal{R}}| = |\mathbb{E}[(c(S)-\eta(S))\mathbf{1}_A]|/\pi_\tau \leq \mathbb{E}[|c(S)-\eta(S)|\mathbf{1}_A]/\pi_\tau \leq \mathrm{CE}_1/\pi_\tau$. The targeting claim follows by applying this at $\tau^\star$ and using $\widehat{\mathcal{R}}(\tau^\star) \leq r$ (right continuity of $\widehat{\mathcal{R}}$ at the infimum, which holds since $c(S)$ has a right-continuous distribution function). If $c = \eta$ a.s. then $\mathcal{R} = \widehat{\mathcal{R}}$, and the per-item form follows because accepted queries have $\eta(S) = c(S) \geq \tau$, so conditional error is $1 - \eta(S) \leq 1 - \tau = r$. The escalation-rate clause holds because $\mathbb{P}(c(S) < \tau)$ involves no labels.

(b) By Theorem 1 with $K = 2$, c and $g \circ c$ have identical achievable sets. For the risk identity, $\{g(c(S)) \geq g(\tau)\} = \{c(S) \geq \tau\}$ by strict monotonicity, so conditioning on the two events is the same and the true risks agree. For non-identification, fix τ_0 interior and r; choose strictly increasing g with $g(\tau_0) = \tau_0$ shaped so that $\mathbb{E}[1 - g(c(S)) \mid c(S) \geq \tau_0] = r$, which is possible because that expectation is continuous in g and ranges over $(0, 1 - \tau_0]$ as g varies over strictly increasing maps fixing τ_0. Then the label-free rule selects τ_0 under $g \circ c$ regardless of the true risk there. □

PROOF OF PROPOSITION 12 (TRANSFER)

(i) On workload W the conditional error of an accepted query with score s is $1 - \eta_W(s) = 1 - c(s) \leq 1 - \tau = r$, since acceptance requires $c(s) \geq \tau$. No property of the marginal is used. (ii) Both π_τ^W and $\mathcal{R}^W(\tau)$ are functionals of the law of $c(S)$ under W. For the TV bound write $\mathcal{R}^W(\tau) = \mathbb{E}_W[(1 - c(S))\mathbf{1}\{c(S) \geq \tau\}]/\pi_\tau^W$, note the integrand is bounded in $[0, 1]$, and bound the differences of numerators and denominators by TV, using $|u_A/v_A - u_B/v_B| \leq (|u_A - u_B| + |v_A - v_B|)/\min(v_A, v_B)$ valid for $u_W \leq v_W \leq 1$. For "can differ arbitrarily", take $c(s) = s$ with $\mathcal{L}_A(S) = \mathrm{Beta}(2, 5)$ and $\mathcal{L}_B(S) = \mathrm{Beta}(5, 2)$: both are calibrated by construction, yet at $\tau = 0.7$

the accept rates are ≈ 0.011 and ≈ 0.580 and the aggregate accepted risks ≈ 0.246 and ≈ 0.174, while both respect the conditional bound ≤ 0.3 from (i). □

PROOF OF THEOREM 13 (FUSION AND CONDITIONAL RECALIBRATION)

(i) Lemma 6 with $\mathcal{G} = \sigma(S) \subseteq \mathcal{H} = \sigma(S, U)$, together with $Q_S(\alpha) = Q^\star(\sigma(S); \alpha)$ under MON from Proposition 7(i).

(ii) Apply the equality clause of Lemma 6 to the same pair: the ceilings agree for all α iff $\mathbb{E}[\Delta \mid \sigma(S, U)] = \mathbb{E}[\Delta \mid \sigma(S)]$ a.s. The left side is W by the tower property, the right side $\Delta(S)$. Negating gives strict inequality for some α iff $\mathbb{P}(W \neq \Delta(S)) > 0$. For the reduction: if $\mathbb{E}[Y_2 \mid S, U] = a_2$ is constant then $W = \mathbb{P}(Y_1 = 1 \mid S, U) - a_2$, so W is $\sigma(S)$-measurable iff $\mathbb{P}(Y_1 = 1 \mid S, U)$ is.

(iii) The attainment clause of Lemma 4 with $\mathcal{H} = \sigma(S, U)$, whose optimizer thresholds $\mathbb{E}[\Delta \mid S, U] = W$; invariance to strictly increasing reparametrization of W is Theorem 1.

(iv) Thresholding $g_Z(S)$ at τ accepts $\{(s, z) : g_z(s) > \tau\} = \{(s, z) : s > g_z^{-1}(\tau)\}$, a group-dependent threshold on s; this is $\sigma(S, Z)$-measurable but not of the form $\{s > \text{const}\}$ unless $g_z^{-1}(\tau)$ is constant in z, so as a function of the pooled score it is non-monotone. Apply (ii) with $U = Z$. The final clause is immediate from Assumption 3: a policy that reads Z is not in $\Phi(S)$. □

Remark 14 (a worked separation for Theorem 13). The gap can be large and can exceed the convexified pooled frontier. Let $Z \in \{A, B\}$ be equiprobable, $S = 0.4$ on group A with $\eta = 0.9$, $S = 0.6$ on group B with $\eta = 0.2$, and tier 2 perfect. At coverage $\alpha = 1/2$: thresholding the pooled S achieves $3/5$, mixing two pooled operating points achieves $31/40$, and the per-group rule achieves $19/20$. Here the pooled score is *anti*-informative relative to the group, which is the regime in which conditional recalibration pays. This mechanism is unavailable to any recalibration of the first score. In our data the analogous fusion gain is real in AUROC and not resolvable on the frontier (§6.2), so this example is a statement about what is possible, not a description of our measurements.

D NUMERICAL VERIFICATION OF THE THEORY

Every counterexample and inequality in §4 was checked before being written, in exact rational arithmetic where the claim is exact.

- **Proposition 7.** Tie-breaking band $[5/6,\ 29/30]$ with the raw score at $29/30$ and the merged score at $9/10$ — exact fractions. The raw threshold policy sits at the *upper* edge of the band, not the middle; mixtures of merged operating points do not recover it (also $9/10$).
- **Proposition 8 under MON.** Maximum over all six reorderings of the three score levels, minus raw, is **exactly** 0: no reordering beats raw when Δ is ordered.
- **Proposition 9.** $5/6 \to 29/30$ exactly, matching the $\sigma(S)$ ceiling from Lemma 4.
- **Theorem 2(ii).** Minimum over $p \in \{0.3, 0.5, 0.7, 0.9\}$, three ROC pairs and 199 coverages of [dominating − dominated] accepted-correct mass is $+1.6 \times 10^{-4} \geq 0$.
- **Theorem 11(a).** Bound verified at $\tau \in \{0.3, 0.5, 0.7, 0.9\}$ on 4×10^6 samples. At $\tau = 0.9$ the accept rate is 0.028 and the bound is 2.39 — vacuous, correctly so, which is why the $1/\pi_\tau$ factor cannot be dropped.
- **Proposition 12.** $\mathrm{Beta}(2, 5)$ versus $\mathrm{Beta}(5, 2)$, both calibrated by construction, $\tau = 0.7$: accept rates 0.011 versus 0.580 (a $53\times$ difference) and aggregate risks 0.246 versus 0.174, while both respect the conditional bound ≤ 0.3.
- **Theorem 13.** Separation of Remark 14: pooled $3/5$, pooled-plus-mixtures $31/40$, per-group $19/20$ — exact.
- **Lemma 4 concavity.** Maximum second difference of $\alpha \mapsto \alpha\,\mathrm{CVaR}_\alpha$ is $+7.5 \times 10^{-6}$, consistent with concavity at grid resolution.
- **Corollary 16.** DKW band recomputed at the analysis-set size $m = 701$, $\delta = 0.05$: 0.051 in coverage, 0.112 in quality after a union bound over the two sub-distributions.

E EXPANDED DISCUSSION OF THE COARSENING AND NON-MONOTONE ROWS

Isotonic regression and histogram binning map distinct scores to a common value, so boundary ties must be broken by something other than the score and the achievable set is undefined until a tie-breaking rule is fixed. Proposition 7 shows $Q_{g(S)}(\alpha) \le Q_S(\alpha)$ for every α and nondecreasing g (equality iff Δ is a.s. block-constant), and that for any label-independent tie-breaker the realized quality lies in a band of width at most the boundary block's probability times the oscillation of Δ on it, whose expectation over the tie-break is again $\le Q_S(\alpha)$ — no gain. Histogram binning is not even weakly monotone, which might suggest Theorem 1 does not apply and the map can help. It cannot, for a different reason — what matters is measurability, not monotonicity: under MON, $Q_{h(S)}(\alpha) \le Q_S(\alpha)$ for *any* measurable h and every α, and likewise for randomized $h(S, V)$ with $V \perp (S, Y_1, Y_2)$ (Proposition 8). Dropping MON opens the one loophole (Proposition 9): with S uniform on $\{0.1, 0.5, 0.9\}$, tier 2 perfect and $\eta = (0.1, 0.9, 0.5)$, at $\alpha = 1/3$ thresholding S achieves $5/6$ while ranking 0.5 above 0.9 achieves $29/30$, the $\sigma(S)$ ceiling — but that is exactly a gain in *discrimination* in the sense of Theorem 2. Even in the loophole what helps is re-ranking, not rescaling; and the *sign* of an observed effect does not separate the two cases, since a noise-driven movement can be positive.

Corollary 15 (the noise band at our sample size)**.** *Fix h and estimate the frontier from m i.i.d. points. The sets $\{h(s) \ge t\}$ have VC dimension 1 in t, so Dvoretzky–Kiefer–Wolfowitz applied to $\mathbb{P}(h(S) \ge t)$ and to $\mathbb{P}(h(S) \ge t, Y_1 = 1)$ gives a uniform deviation of at most $\sqrt{\log(2/\delta)/2m}$ in each with probability $\ge 1 - \delta$: at $m = 701$, $\delta = 0.05$, that is 0.051 in coverage and, after a union bound, 0.112 in quality — an order of magnitude above the movements we measure for isotonic (1.83×10^{-2}) and histogram binning (2.28×10^{-2}).*

F EXPANDED DISCUSSION OF THE CALIBRATION-UTILITY RESULTS

These theorems do not say calibration is useless; they locate its role. Calibration does not select the frontier, it selects a *parametrization* of it — and a parametrization is what is needed to hit a target without labels. With c the reported confidence, $\mathcal{R}(\tau) := \mathbb{E}[1 - Y_1 \mid c(S) \ge \tau]$ the true accepted risk, $\widehat{\mathcal{R}}(\tau) := \mathbb{E}[1 - c(S) \mid c(S) \ge \tau]$ its label-free surrogate and $\pi_\tau := \mathbb{P}(c(S) \ge \tau)$, Theorem 11 bounds $|\mathcal{R} - \widehat{\mathcal{R}}| \le \mathrm{CE}_1/\pi_\tau$, so the label-free choice $\tau^\star := \inf\{\tau : \widehat{\mathcal{R}}(\tau) \le r\}$ realizes $\mathcal{R}(\tau^\star) \le r + \mathrm{CE}_1/\pi_{\tau^\star}$, and under exact calibration the per-item guarantee is simply $\tau = 1 - r$. The $1/\pi_\tau$ factor is not optional — conditioning on acceptance re-weights calibration error onto the accepted region — and it predicts a fact we measure: the bound goes vacuous exactly where the accept set collapses (§6.4). Part (b) of the same theorem shows why calibration is the right device and nothing else is: for strictly increasing g, c and $g \circ c$ induce the same frontier but $\mathcal{R}_{g\circ c}(g(\tau)) = \mathcal{R}_c(\tau)$, so the risk attached to a given numerical threshold varies arbitrarily with g and $\tau \mapsto \mathcal{R}(\tau)$ is *not identified* from the unlabeled score distribution. Calibration is precisely the choice of scale making $\widehat{\mathcal{R}}$ agree with $\mathcal{R}$. Transfer splits the same way (Proposition 12): the *per-item* guarantee transfers exactly, with no assumption on marginals, while the accept rate and *aggregate* accepted risk do not, since both are functionals of the law of $c(S)$ — two Beta workloads calibrated by construction give accept rates 0.011 and 0.580 at the same $\tau = 0.7$ while both respect the conditional bound. A calibrated cascade can promise "every answer we keep is at least $1 - r$ likely correct" on a new workload, but not "we will escalate 20% of it".

What *can* move the frontier is an enlargement of the σ-algebra. With U any auxiliary signal and $W := \mathbb{E}[Y_1 - Y_2 \mid S, U]$, Theorem 13 gives $Q_{S,U}(\alpha) \ge Q_S(\alpha)$ with equality for *all* α iff $W = \Delta(S)$ a.s.: fusion strictly improves the frontier somewhere iff U says something about the escalation value conditional on S, and the gain is realized by thresholding W. The same statement covers per-group recalibration, where the pooled score $g_Z(S)$ is non-monotone in S and the policy is $\sigma(S, Z)$-measurable — provided Z is observed at inference time; if not, the policy is inadmissible under (A3). The contrast with Proposition 8 is the one to check: both concern non-monotone transformations and reach opposite conclusions, because histogram binning stays inside $\sigma(S)$ while per-group recalibration does not. The prescription is a conditional-independence test — does U move $\mathbb{P}(\text{correct} \mid s, u)$ away from $\mathbb{P}(\text{correct} \mid s)$ — not a calibration metric. The theorem says only that the gain is *available*; §6.2 shows that in our data it is real in AUROC and not resolvable on the frontier. On scope, Theorem 1 holds for arbitrary K; the ceiling and dominance results are proved for a fixed

tier pair and extend by backward recursion — but Δ then depends on the downstream policy, so MON must be re-checked at each stage, and we do not claim the general-K ceiling results.

G PRICE SENSITIVITY AND CASCADE ECONOMICS IN FULL

Expanding §6.4: in the $k = 1$ regime, 0 of 21 swept points at or under opus-only cost reach opus-only accuracy, because the haiku pass is paid on every query whether or not it is accepted, so full escalation costs \$0.01137 against opus-only's \$0.00930. For the three-tier comparison the 0.87–0.88 band advantage is $+8.4\%$ at the median, and the mechanism is repair-per-dollar: per escalated query sonnet costs $0.64\times$ opus but repairs only $0.79\times$ as many tier-1 errors (0.411 vs 0.518) while doing $1.8\times$ more damage to items tier 1 had right.

Price moves the optimal rate, not the invariance. Sweeping the opus:haiku price ratio from $2\times$ to $50\times$, the largest frontier-AUC deviation induced by any strictly monotone calibrator is 1.2×10^{-3}, attained on `haiku_agree` ($k = 5$); the per-signal maxima are 5.7×10^{-5} (`haiku_verb`, $k = 5$), 5.0×10^{-4} (`haiku_verb1`, $k = 1$) and 1.2×10^{-3} (`haiku_agree`, $k = 5$), so the *global* strict-monotone maximum is 1.2×10^{-3} and the headline signal's 5.7×10^{-5} is its minimum, not the overall bound. All three sit below the $1/701 = 1.43 \times 10^{-3}$ frontier resolution of a single item. Weakly monotone calibrators reach 4.5×10^{-3}, from ties. These residuals are not the exact zeros of Tables 2–3. We have not isolated what produces them — the candidates are this sweep's separate frontier-AUC evaluation path and the same negative-slope and float64-saturation effects diagnosed in Appendix O — and we state that rather than assert a mechanism we did not test. What we *did* verify is the decisive point: each residual is *numerically identical at all ten price ratios* — `haiku_verb`'s 5.7×10^{-5} at $2\times$ and at $50\times$ alike, and likewise for each other signal at its own value — so whatever produces it is a property of the evaluation and carries no price dependence, consistent with the fact that a monotone map changes only the labels on the score while price enters the frontier only through the cost axis. The *optimal escalation rate*, by contrast, is entirely price-dependent: under a utility criterion of \$0.20 per correct answer with a single-sample first tier, it runs 0.489 at $2\times$, 0.170 at $5\times$, 0.140 at $10\times$, 0.050 at $20\times$ and 0.000 at $30\times$ and above. Any claim of the form "escalate about $X\%$ of queries" is a claim about a price, not about a method.

Self-consistency does not pay for itself. Eliciting five haiku samples costs \$0.00979 per query — 105% of the price of answering every query with opus (\$0.00930) — so a cascade whose gate needs five tier-1 samples is cost-dominated before it routes anything. The $k = 1$ regime (\$0.00207, 22.3% of all-opus) is the only economically live one, and first-sample verbalized confidence matches the five-sample signals in AUROC (0.761 vs 0.767, fully overlapping intervals): the cheap signal is as good as the expensive one, fortunately, since the expensive one costs more than the answer it is trying to avoid buying. In that regime the cascade is genuinely useful *below* opus accuracy (0.8973 at \$0.00816, 88% of opus's price for 0.4 pp less accuracy), but 0 of 21 swept points at or under opus-only cost reach opus-only accuracy — matching it requires escalating everything, at \$0.01137, because the haiku pass is also paid for. The middle tier likewise does not earn its place: with sonnet inserted between haiku and opus the three-tier frontier is indistinguishable from or worse than two-tier at every accuracy target above 0.88 under every refusal convention, and in the 0.87–0.88 band the apparent 10–15% cost advantage has an undetermined sign ($+8.4\%$ median, 95% CI $[-49.0\%, +39.4\%]$ at 0.88). The mechanism is repair-per-dollar: per escalated query sonnet costs $0.64\times$ opus but repairs only $0.79\times$ as many tier-1 errors (0.411 vs 0.518) while doing $1.8\times$ more damage to items tier 1 had right. We report this as an unresolved sign, not as a finding that middle tiers never help.

Why matching opus-only accuracy requires escalating everything. At the measured $4.5\times$ price ratio the haiku pass is paid on every query whether or not it is accepted, so the cascade's cost at full escalation is \$0.01137 against opus-only's \$0.00930. There is therefore no operating point at or below opus-only cost that reaches opus-only accuracy; the cascade's value in the $k = 1$ regime is strictly in the sub-opus-accuracy region. This is a property of the price ratio, not of the policy, which is why §6.4 sweeps it.

H STRUCTURAL LIMITS, PER-CELL

Three limits are properties of the data rather than the methods, and reporting numbers past them would be reporting artifacts. **Some targets are unattainable, not merely hard.** Requiring $\geq 10\%$ accepted, the minimum attainable risk even for an oracle threshold is 0.031 for haiku verbalized confidence pooled (0.073 for haiku self-consistency, 0.066 for sonnet verbalized), so $r = 0.02$ **is unattainable pooled for every signal and calibrator**, flagged unattainable in 16 of 24 cells. There the accept set collapses: median accepted test items is 1 for temperature and Platt against 142 for raw, and the set is *empty* in 27–53% of replicates. The apparent "excellent" realized risks (0.001 temperature, 0.004 Platt) are computed over a handful of items and **must not be quoted as successes** — exactly as Theorem 11(a) predicts, since the $1/\pi_\tau$ factor makes the bound vacuous as the accept probability vanishes. **Some targets are non-binding:** any r above the tier's marginal error rate (0.144 haiku, 0.130 sonnet, pooled) is satisfied by accepting everything, so pooled $r \geq 0.15$ is uninformative; we flag `binding` and exclude those rows from every headline number. **Per-task resolution:** with $n = 219$–249 per task and 100–200 accepted items at binding targets, no per-task risk difference below about 0.03 is resolvable.

Corollary 16 (the noise band at our sample size)**.** *Fix h and estimate the frontier from m i.i.d. points. The sets $\{h(s) \geq t\}$ have VC dimension 1 in t, so Dvoretzky–Kiefer–Wolfowitz applied to $\mathbb{P}(h(S) \geq t)$ and to $\mathbb{P}(h(S) \geq t, Y_1 = 1)$ gives a uniform deviation of at most $\sqrt{\log(2/\delta)/2m}$ in each with probability $\geq 1 - \delta$: at $m = 701$, $\delta = 0.05$, that is 0.051 in coverage and, after a union bound, 0.112 in quality. If h is* fitted *on the same data the class is unions of up to B intervals and the band is larger.*

I EXPANDED POSITIONING AGAINST THE TWO CLOSEST PAPERS

Two 2026 papers place calibration at the centre of the escalation decision, UCCI (Kotte, 2026) and Conformal Cascade (Dou et al., 2026), and both deserve a fuller treatment than main-text space allows.

Mapping UCCI's assumptions onto Theorem 2. UCCI's cost-optimality result (their Theorem 1) is stated under three explicit assumptions: (i) the large model costs more than the small model; (ii) the large model's accuracy on escalated queries equals its marginal accuracy on the full test distribution, independent of which queries are escalated; and (iii) the router has access to a calibrated error probability $\hat{p}(x) = \mathbb{P}(\text{error} \mid u(x))$, where $u(x)$ is their token-margin uncertainty score. Assumption (ii) is, term for term, our constant tier-2 accuracy condition in Theorem 2 ($a_2(s) \equiv a_2$): both results need the escalated tier's error rate to not depend on which queries reach it. Assumption (i) is simply our cost ordering $c_1 < c_2$. Assumption (iii) is the one our theorem renders unnecessary. Theorem 2 needs only that the router's score induce some ROC curve R against the base rate p; it does not need that score to be a calibrated probability. UCCI's own paper checks assumption (ii) directly on its workload, finding the large model's accuracy on escalated queries (0.928) close to its accuracy on the full test set (0.932), a gap of 0.004. That is exactly the check our Theorem 2 also needs, and it is a check about the fixed-tier-accuracy structure of the problem, not about calibration.

What UCCI's own results are consistent with is that the frontier gain in their setting comes predominantly from the token-margin signal, not from the calibration step. Their paper reports a baseline, labeled "conformal prediction," that uses the identical raw uncertainty score $u(x)$ as a nonconformity score and selects a threshold by distribution-free coverage control rather than by calibrating $u(x)$ to a cost-minimizing probability. At the same accuracy target, this uncalibrated-signal baseline reaches cost 2.18 against UCCI's 2.08: a gap of under 5%, holding the signal fixed and varying only whether the threshold-selection procedure passes through a calibrated probability. By contrast, their entropy-thresholding baseline, which changes the signal rather than the calibration step, is 11% more costly than UCCI. On this evidence the larger share of UCCI's advantage over a naive baseline traces to using token-margin uncertainty at all, not to calibrating it; the calibration step's marginal contribution, isolated by the conformal-prediction comparison, is the smaller of the two effects. We did not run this comparison ourselves; it is UCCI's own ablation, and we read it as supporting evidence for our attribution claim rather than as something we have independently verified.

Conformal Cascade. The second paper stands in a different relationship to our results: complement rather than contrast. Conformal Cascade (Dou et al., 2026) uses conformal prediction set size as the deferral rule for a distribution-free accuracy guarantee across tiers. Conformal methods (Angelopoulos & Bates, 2021) deliver a distribution-free guarantee by construction, and prediction-set size is a coarsening of the score, row (b) of Table 1, so our theory predicts it is frontier-neutral-or-worse relative to thresholding the underlying score continuously, while buying a guarantee a raw threshold does not have. That is the trade Theorem 11 and Proposition 12 formalize on the calibration side, and §6.4 sharpens which promise survives a workload shift: the per-item risk bound, not the accept rate.

J THE TIE-BREAKING ENVELOPE FOR COARSENING CALIBRATORS

The coarsening calibrators do deviate — up to 1.83×10^{-2} (isotonic) and 2.28×10^{-2} (histogram) — as Proposition 7 predicts, since both pool distinct scores into blocks and the policy must then break boundary ties by something other than the score. Tested against the best/worst tie-breaking envelope for every (signal, calibrator, rate) cell, **raw accuracy lies inside the envelope in 69 of 70 cells and calibrated accuracy in 70 of 70**: the movement is a coin flip's width, not information, and an order of magnitude inside the $m = 701$ noise band.

K THE REPAIR-RATE BOUND ON FRONTIER GAINS

The damping has an arithmetic explanation, and it is the most transferable number we report. Escalation can only help where the strong tier repairs the weak one, and $\mathbb{P}(\text{opus correct} \mid \text{haiku majority wrong}) = 0.4554$ ($n = 101$) against $\mathbb{P}(\text{opus wrong} \mid \text{haiku majority right}) = 0.0233$ ($n = 600$), so an escalation slot is worth at most 0.455 accuracy points, not 1: raising escalated-set precision by δ at rate q buys about $q\,\delta$ $(0.455{+}0.023)$, and fusion's measured $+0.071$ precision gain at 20% escalation therefore predicts $+0.68$ pp against a measured $+0.34$ pp — even the *predicted* effect below the noise floor at this n. **Any paper claiming a large frontier gain from better confidence scoring on a task pair with a sub-50% repair rate is claiming something arithmetic forbids**; we recommend cascade papers report this rate, since it bounds every policy result they contain. Ranking closes $60.7/48.8/43.8/41.4/52.0\%$ of the random-to-oracle gap at $10/20/30/40/50\%$ escalation ($k = 5$) and fused out-of-fold $53.6/53.7/46.9/48.3/56.0\%$ — roughly half the available work, with the unclosed half largely where escalation cannot help at all.

L RISK-TARGETING RESULTS IN FULL

Aggregating all 24 (tier × signal × scope) cells at attainable binding targets, mean absolute gap between realized and target risk is 0.046 (raw), 0.048 (temperature), 0.057 (Platt), 0.044 (isotonic), 0.057 (histogram): no calibrator improves absolute tracking in aggregate, and the ranking among them is unstable across cells. Per-cell paired deltas against raw are $+0.0066$ $[+0.0058, +0.0074]$ (temperature), $+0.0050$ (Platt), $+0.0011$ (isotonic) and $+0.0057$ (histogram) on the haiku verbalized-confidence pooled cell. Full per-cell values, including the `binding` and `target_attainable` flags used to select rows, are in `targeting.csv`.

M TRANSFER, PRICE, AND ECONOMICS: THE FULL STATEMENTS

Three further results, with full tables in Appendices N–G. *Transfer* shows the asymmetry Proposition 12 predicts: applying a τ fitted on task A to task B over all six ordered pairs, mean |realized risk − target| on B is 0.049–0.086 while the mean accept-rate shift is 0.104–0.174, a $3\times$ larger mismatch (a temperature map fitted on MATH-500 accepts 1.4% of MATH-500 items but 57.0% of GSM8K items), so **a practitioner sizing their strong-tier budget from the calibration workload's accept rate can be off by a factor of two or more** — though "the risk promise survives" means a mismatch of the same order as the targets themselves. *Price* separates two easily conflated claims: over a $2\times$–$50\times$ opus:haiku sweep the largest frontier-AUC deviation from any strictly monotone calibrator is 1.2×10^{-3} (per-signal maxima 5.7×10^{-5} to 1.2×10^{-3}, all below the $1/701$ single-item frontier resolution) — not the exact zero of Tables 2–3, because the sweep re-evaluates on a coarser threshold grid, but *numerically identical at all ten price ratios* and so carrying no price dependence — while the

cost-optimal escalation rate moves from 0.489 at $2\times$ to 0.000 at $30\times$. Any claim of the form "escalate about $X\%$ of queries" is a claim about a price, not a method. *The gate can cost more than the answer*: five haiku samples cost \$0.00979 per query, 105% of answering everything with opus (\$0.00930), so the $k = 1$ regime (\$0.00207, 22.3% of all-opus) is the only economically live one — and first-sample verbalized confidence matches the five-sample signals in AUROC, so the cheap signal is as good as the expensive one. There the cascade is useful *below* opus accuracy (0.8973 at \$0.00816) but never at it, and the middle tier does not earn its place either: three-tier is indistinguishable from or worse than two-tier above 0.88 accuracy under every refusal convention, and in the 0.87–0.88 band the apparent cost advantage has an undetermined sign (95% CI $[-49.0\%, +39.4\%]$).

N TRANSFER ACROSS WORKLOADS, IN FULL

Transfer. Proposition 12 predicts a specific asymmetry, and it is what we find. Fitting calibration on task A and applying the same τ to task B across all six ordered task pairs at binding targets, the per-item risk guarantee survives while the accept rate does not: mean $|\text{realized risk} - \text{target}|$ on B is 0.049–0.086 depending on calibrator, while the mean accept-rate shift is 0.104–0.174 — roughly a $3\times$ larger mismatch, maximum 0.697. Concretely, a temperature map fitted on MATH-500 accepts 1.4% of MATH-500 items but 57.0% of GSM8K items. **A practitioner who sizes their strong-tier budget from the accept rate observed on the calibration workload can be off by a factor of two or more**, even though the risk promise roughly survives — where "roughly" means an absolute mismatch of the same order as the targets themselves, so this is a relative statement rather than a tight guarantee, and 147 of 570 cross-task rows are flagged thin-accept in `transfer.csv`.

O DIAGNOSTICS FOR THE TWO RESIDUAL NONZERO DEVIATIONS

§6.4 reports that strictly monotone calibrators leave the realized escalation rate within 0.0013 (temperature) and 0.0210 (Platt) of the raw score's, rather than at exactly zero. Both residuals are diagnosed and neither contradicts Theorem 1.

Negative fitted Platt slopes. A regularized Platt fit lands on a *negative* slope in a minority of bootstrap resamples on low-cardinality or near-separable signals — up to 20.8% of resamples for sonnet self-consistency on GSM8K. A decreasing map is outside Theorem 1's hypothesis (it converts up-sets into down-sets, equivalent to thresholding $-s$), and conditioning on the fitted slope separates the two populations completely: on haiku verbalized confidence/GSM8K, escalation-set disagreement against the raw score is 0.00011 over the increasing fits and 0.359 over the decreasing ones. This is why we report the fitted slope for every fit rather than assuming its sign.

Float64 saturation of shallow increasing maps. When an increasing map has a shallow slope, distinct inputs can round to the same double. In 27 of 300 resamples on one cell an increasing fit with $A = 0.233$ compressed 26 distinct scores to 25, creating a tie that did not exist in the raw score and hence a boundary ambiguity of exactly the kind Proposition 7 describes. This accounts for the entire residual 0.0013 observed for temperature scaling; it is a floating-point artifact of the implementation, not a property of the map.

P ELICITATION PROTOCOL DETAILS

Sampling. Each item is presented to each tier with an identical prompt (143 tokens; identical across tiers by construction, which is what makes the cost arithmetic of §3 exact). Haiku is sampled $k = 5$ times, sonnet $k = 3$, opus $k = 1$, all at temperature 1.0. Verbalized confidence is requested in-prompt alongside the answer and parsed per sample; self-consistency signals are computed over the sample set after exact-match normalization of answers.

Signal definitions. `verb_mean`, `verb_min`, `verb_first` and `verbstd` are the mean, minimum, first-sample value and standard deviation of parsed verbalized confidence over the k samples. `agree` is the modal-answer share; `entropy` is the negated Shannon entropy of the answer distribution; `n_distinct` is the negated count of distinct answers. All are oriented so that higher means more confident, and mapped to $[0, 1]$ before calibration.

Degenerate-fit handling, reported rather than hidden. Histogram binning degenerates on low-cardinality signals: self-consistency over three sonnet samples takes four distinct values, so requesting ten quantile bins collapses the edges. We reduce the bin count to the signal's cardinality and fall back to the raw score when fewer than three distinct values remain, recording the count of such fits in the released tables (up to 406 fits in one cell). Histogram results on self-consistency and negated entropy should be treated as unreliable for that reason. Platt fits with a negative fitted slope are recorded rather than discarded, and §6.4 conditions on the slope; a decreasing map is outside Theorem 1's hypothesis and we do not claim the invariance for it.

What we did not measure. No logit-derived signals (max-probability, token margin, sequence likelihood) — our access is to text output only. No opus-side self-consistency, since opus is $k = 1$; opus's own confidence signals are weak in any case (verbalized $\mathrm{AUROC}\ 0.664$, self-consistency 0.522, the latter necessarily uninformative at $k = 1$). All targeting results are for the escalating tiers, which is where the decision lives.

A property of the signal worth knowing. Verbalized confidence is heavily concentrated near 1.0 (median 0.964 for haiku, 5th percentile 0.72). This is why absolute-threshold methods are fragile at tight targets, and why the discrimination floor of §6.4 sits where it does. It is a property of the model's confidence reports, not of the calibrators.