CREDO: A CERTIFIED VLM-POOL ROUTER BOUNDS FALSE ACCEPTS BUT NOT THE ERROR ITS USERS RECEIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

A per-tier false-accept bound does not bound the error of the traffic it admits; that error is set by the base rate and by how well the probe separates right answers from wrong ones. We show this on CREDO, a training-free router over a pool of vision-language models (VLMs) and, to our knowledge, the first to certify a user-set answer-quality budget at every operating point of such a pool: a finite-sample, distribution-free, per-tier bound at the user's budget α on the rate at which a tier's wrong answers are accepted, built from a rule-based probe, split-conformal thresholds and a cheapest-certified-first cascade. It covers the queries clearing those thresholds; under the default strongest-tier fallback the rest ship uncertified. On a frozen four-tier pool over six benchmarks and 9 calibration splits, the bound holds on average for the quantity it bounds and fails for the quantity a deployer reads: the conditional error of certified traffic exceeds α at 23 of 30 operating points, stable across splits. The calibration fold predicts the gap before it is measured, a label audit leaves it essentially unchanged, no weak-signal channel raises pooled discrimination appreciably on any split under the shipped pooling, and a learned fusion of the same inputs and a hidden-state probe above the registered bar still leave most operating points over budget. Routing adds no accuracy over the best fixed tier and, once the probe is charged, no cost saving over the strongest tier. The certificate's reach is real and its reach grows with the risk level, but at every level the error a deployer reads is set by the base rate and the probe rather than by the bound.

1 INTRODUCTION

Open-weight vision-language models (VLMs) span two orders of magnitude in serving cost, and a stack serving several must pick one per request. Routing benchmarks (Huang et al., 2025; Ma et al., 2026) and learned routers (Cheng & Dong, 2026; Tang et al., 2025; Šabanović et al., 2026) report average cost-accuracy operating points, and none states, for the point a deployer selected,

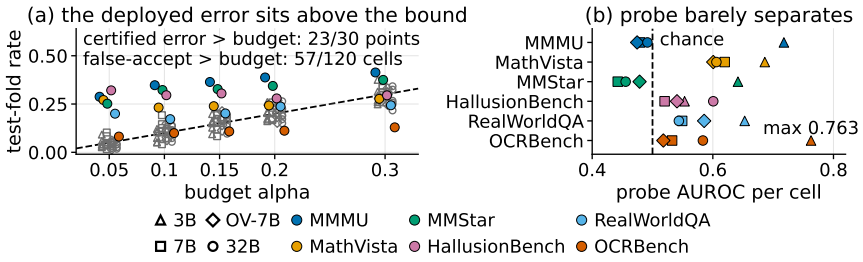

Figure 1: **What the certificate bounds, what a deployer reads, and the probe behind the gap.** (a) The per-tier false-accept rate Eq. 1 bounds (gray) straddles the budget line (dashed); the conditional error of certified traffic (colour) sits above it. (b) Probe AUROC per (benchmark, tier) cell: most cells sit near chance; the one benchmark meeting its budget at $\alpha \geq 0.1$ (OCRBench) holds the most separable cell.

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.

that the risk of a wrong answer stays inside a budget. Guarantee-bearing routers are text-only, and most act on pairs of models (Su et al., 2025; Hao et al., 2026); the closest, LEC (Wang et al., 2025a), controls selection-conditioned risk in text routing, the quantity audited here; the conformal toolbox (Angelopoulos & Bates, 2021; Angelopoulos et al., 2024) certifies one predictor, not an allocation over a cost-ranked pool. We build that certificate and audit what it does and does not bound.

CREDO (Certified Risk-controlled routing over VLM pools) attaches that certificate; whether it closes the gap is what we audit. Every operating point carries a finite-sample, distribution-free certificate: a per-tier bound on the rate at which a tier's wrong answers are accepted (Figure 2). A training-free rule-based probe fuses weak difficulty signals from the smallest model's draft forward pass (logprob and entropy statistics, a k-sample vote entropy, question-structure features, CLIP image-text alignment (Radford et al., 2021)) into a per-tier can-answer score. Split-conformal calibration on a held-out fold turns each score into a threshold $\tau_m(\alpha)$ whose false-accept rate is bounded at the budget α; this is label-conditional split conformal on the wrong stratum (Angelopoulos & Bates, 2021), one threshold per tier, not CP-Router's prediction-set rule (Su et al., 2025), which needs option probabilities. The router walks the pool cheapest-first to the first tier clearing its threshold, and a mini-max rule escalates or abstains when none does.

We evaluate on a frozen four-tier pool (Qwen2.5-VL-3B/7B/32B, LLaVA-OneVision-7B (Bai et al., 2025; Li et al., 2024)) over MMMU, MathVista, MMStar, HallusionBench, RealWorldQA, and OCRBench (Yue et al., 2024b; Lu et al., 2024; Chen et al., 2024a; Guan et al., 2024; xAI, 2024; Liu et al., 2024), 6,294 questions and 25,176 generations, with baselines re-run in-pool and every key quantity re-derived under 9 split seeds. The certificate holds on average over them for the quantity it bounds and fails for the quantity a deployer reads: the conditional error of certified traffic exceeds α at 23 of 30 operating points (23–25 across splits). Routing does not add accuracy either: against the best fixed tier of each test fold, chosen with test labels and so an oracle comparator, no point exceeds it by more than +1.6 points at any of 270 (benchmark, budget, split) combinations; learned routers tuned on calibration labels land in the same cost band at similar quality with no risk statement. The gap is not label noise: a three-stage audit leaves the pattern essentially unchanged (24 of 30 before, 23 after). It is not the construction: certifying the accepted region directly would collapse coverage to a median of 0.5% at α=0.10. It is the base rate together with the probe's discrimination, through an identity that predicts the gap before measurement (Section 5): the certified error of the cascade replayed on the calibration fold predicts the test-fold value (r=0.90, against r=0.85 from the base rate alone), its excess over budget falls with each cell's AUROC (Spearman $\rho = -0.94$, a sign the identity implies), no weak-signal channel lifts pooled discrimination beyond an AUROC of 0.572 on any split (0.692 where its strongest channel is defined), a learned fusion of the same inputs reaches 0.679 and leaves 21–23 of 30 points over budget, and a pre-registered hidden-state probe above the signal bar leaves 20 of 30 (the rule probe on the same halved fold, 24), the registered reading, stated at its scope: these three probes at this budget do not remove over-admission.

Contributions: (i) a training-free certified router over VLM pools, a wrong-stratum conformal threshold per tier of a cost-ranked pool (Section 3); (ii) a certified-frontier protocol with in-pool baselines and per-point realized risk, every headline re-derived across 9 splits (Tables 1–2, Figure 4); (iii) an audit of that certificate against the error routed traffic receives, its base-rate mechanism predicted, then measured (Figure 6); a false-accept bound and selective error differ by an identity, so the contribution is the measurement, not the distinction; (iv) a pinned, CPU-reproducible artifact.

2 Related Work

Routing and cascades. Multimodal routing has harnesses (Huang et al., 2025; Ma et al., 2026) and routers (Cheng & Dong, 2026; Tang et al., 2025; Šabanović et al., 2026; Dwivedi et al., 2026); text routing is mature (Ong et al., 2024; Ding et al., 2024; Hu et al., 2024; Li, 2025; Moslem & Kelleher, 2026); cascades (Chen et al., 2023; Aggarwal et al., 2024; Yue et al., 2024a; Valkanas et al., 2025; Hou et al., 2025) stop at the first model whose confidence clears a tuned threshold, as do alarm thresholds outside model selection (Xiao et al., 2025). All optimize average cost-accuracy trade-offs; none bounds answer-quality risk at a chosen operating point.

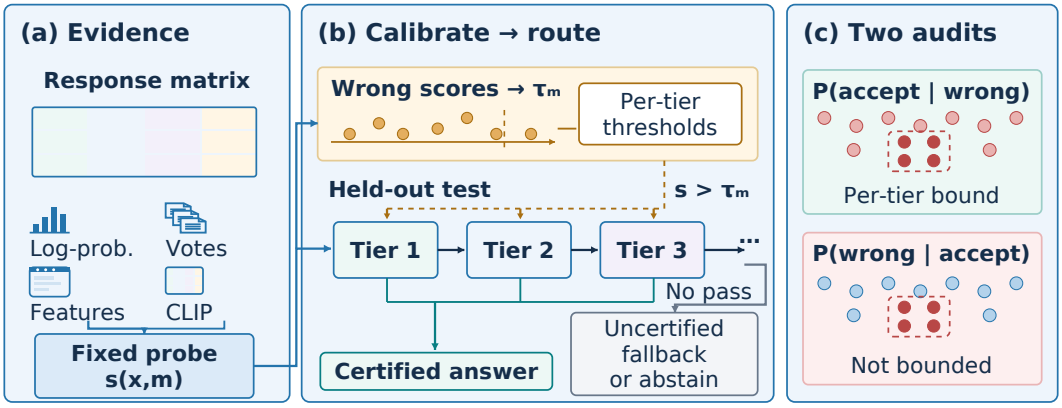

Figure 2: **CREDO.** *(a)* A one-time GPU response matrix and cheap per-query signals feed a fixed rule probe; every downstream score is a CPU function of the cache. *(b)* Wrong-subset calibration sets each tier's threshold τ_m; the cascade certifies the first tier that clears it, else falls back or abstains, uncertified. *(c)* The audit reads one accepted-and-wrong mass under two denominators, $P(\text{accept} \mid \text{wrong})$ and $P(\text{wrong} \mid \text{accept})$. Only calibration and the audit read labels.

Distribution-free risk control and certified selection. Split conformal prediction (Angelopoulos & Bates, 2021) and Conformal Risk Control (Angelopoulos et al., 2024) certify a single predictor and say nothing about allocating a pool against cost. Routers that couple them to selection (Su et al., 2025; Hao et al., 2026; Kotte, 2026; Overman & Bayati, 2025; Wang et al., 2025a; Anonymous; Dou et al., 2026; Guo et al., 2026) are text-only and validated at a few operating points, and multimodal cascades admit forced deferral of the routing decision itself (Liu et al., 2026); vision-side certificates cover judge agreement, image classification and tool-using reasoning (Jung et al., 2024; Turkoglu et al., 2026; Fayyazi & Akrami, 2026), not routing over a cost-ordered VLM pool. The same accept-or-defer construction holds a labeling-risk target by withholding uncertain items (Xiao et al., 2026b), and a further family checks the answer instead, by construction (Xiao et al., 2026a). No prior work we know of certifies a user-set answer-quality budget at every operating point of a VLM pool while remaining training-free; the premise shared by conformal routers (Su et al., 2025; Hao et al., 2026), that a per-model conformal threshold makes a routed answer trustworthy at the stated budget, is what our audit tests (Section 4.3).

3 Method

Problem and certificate. A pool $\mathcal{M} = \{m_1, \ldots, m_K\}$ of frozen VLMs is arranged in a fixed cascade order by design capacity, with c_m the measured GPU-seconds per query (not monotone in that order here: the smallest tier carries the probe forward and is the most expensive, Section 4.1). A query $x = (I, q)$ pairs an image with a question; each pool member's answer has a correctness $y_m(x) \in \{0, 1\}$ decided by the benchmark's official scorer, and a router maps x to one tier or to abstention. The deployer sets a risk budget $\alpha \in (0, 1)$, and the router minimizes expected cost subject to a per-tier certificate on the acceptance rule,

$$P\big(\text{tier } m \text{ accepted} \;\big|\; y_m(x) = 0\big) \;\le\; \alpha, \tag{1}$$

over the test distribution, with finite-sample, distribution-free validity. Eq. 1 caps how many wrong answers leak *through* each certificate; it does not by itself bound the error rate *inside* the accepted region, which also scales with the tier's marginal wrong rate (Figure 3; Section 5 writes the identity down). We audit both quantities on the test fold and report violations rather than patch them; sweeping α traces the *certified cost-quality frontier*. CREDO is a post-hoc layer over a one-time *response matrix*; scoring, calibration, routing and every ablation are CPU-only functions of that cache, and labels (official scorers ported 1:1, Appendix B) appear in the calibration fold and the marked evaluation blocks, never in a routing decision.

Weak-signal probe. Per query CREDO computes cheap difficulty signals from the smallest tier's draft forward pass plus CPU features: (i) per-token logprob mean and truncated top-20 entropy of the 3B tier's draft; (ii) vote entropy of k=5 self-consistency samples from the same tier, a frozen

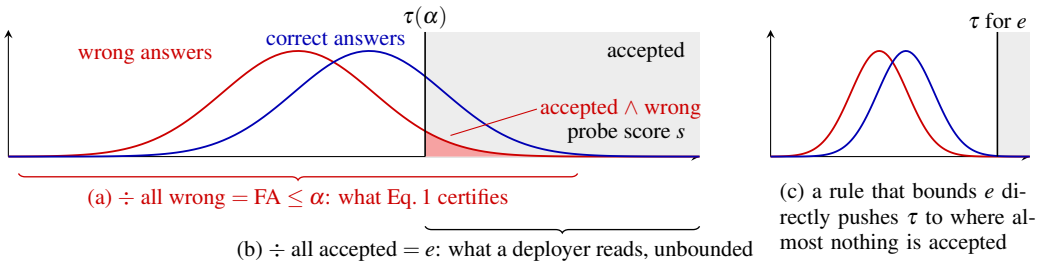

Figure 3: One shaded region, two denominators: the accepted-and-wrong mass over all wrong answers is the false-accept rate Eq. 1 bounds (a); over all accepted answers it is the error a deployer reads (b), which nothing bounds. Bounding (b) directly empties the accepted region (c).

model's agreement with itself read as a difficulty cue rather than as the answer; (iii) question features: length, option count and type, OCR-density and chart/numeric keyword cues, image geometry; and (iv) CLIP image-text alignment (Radford et al., 2021). Fusing such cues is a well-mapped design space (Xu et al., 2026); CREDO takes its simplest point, a rule probe with fixed, documented, unfitted weights that yields a per-tier can-answer score $s(x, m) \in [0, 1]$, a capacity prior rising with cost rank, penalized by difficulty, harder on cheaper tiers. Only the *ordering* of scores enters calibration; a worse probe widens thresholds and defers more traffic but cannot invalidate Eq. 1.

Split-conformal certification. Question ids are split 50/50 into calibration and test folds by a seeded permutation. On the calibration fold, for each tier m CREDO takes the *wrong subset* $\{s(x, m) : y_m(x) = 0\}$ and sets the threshold to the conservative split-conformal $(1-\alpha)$-quantile,

$$\tau_m(\alpha) \;=\; \text{Quantile}\Big(1-\alpha;\; \big\{s(x,m) : y_m(x) = 0\big\}\Big), \tag{2}$$

at the finite-sample rank $\lceil (n+1)(1-\alpha) \rceil$ (Angelopoulos & Bates, 2021). Accepting tier m only when $s(x, m) > \tau_m(\alpha)$ bounds the accepted fraction of its *wrong* answers at α under exchangeability of the two folds, label-conditional split conformal on the wrong stratum (Angelopoulos & Bates, 2021), one threshold per tier; CP-Router (Su et al., 2025) routes by conformal prediction-set size instead and enters only as a probe-score variant (Appendix C). A tier with fewer than $(1-\alpha)/\alpha$ wrong calibration answers gets a vacuous threshold, is flagged, and is never treated as certified. Bounding the error *inside* the accepted region directly (RCPS/conformal risk control (Angelopoulos et al., 2024)) would cover almost no traffic with these probe scores (Section 5), so we keep the false-accept construction and audit its deployment-facing gap.

Cascade, fallback, and sweep. CREDO walks the pool in cascade order to the first tier whose score strictly clears its threshold. When none does, a mini-max rule picks the tier minimizing the worst-case endpoint of its calibration-derived marginal wrong-rate interval (empirical and $+1$-conservative rates), the strongest-model fallback derived from data, or, when the policy allows, abstains, with abstentions counted separately; every delivered answer is certified at α or marked as fallback or abstention. Sweeping α over $\{0.05, 0.10, 0.15, 0.20, 0.30\}$ re-runs calibration and routing per point, audited on the test fold twice (the false-accept rate Eq. 1 bounds; the certified-traffic error it does not), and the pipeline re-runs per split seed in seconds (Appendix E).

4 Experiments

4.1 Setup

The frozen pool is Qwen2.5-VL-3B (probe tier, 3.07 s/item), Qwen2.5-VL-7B (0.52), LLaVA-OneVision-7B (0.66), and Qwen2.5-VL-32B-AWQ (strongest by design, 1.26), served through one vLLM stack with per-item costs smoke-measured there (Table 4, Appendix A). Every router here reads the tier-1 probe (k=5 self-consistency draws with logprobs on a greedy draft), so each routed query, abstentions included, pays that forward (3.07 s/item, the tier-1 rate) plus the rate of any higher tier that answers it; tier 1's answer is the draft already paid for, fixed tiers pay no probe. Cascade order is thus not cost order, and every router-versus-fixed comparison carries the probe as a constant (Appendix E). Six public benchmarks at their official splits and pinned revisions (Table 5) give

6,294 questions and 25,176 generations. The seeded 50/50 split (Section 3) is the only randomness downstream of the matrix, so every headline quantity is re-derived under 8 additional split seeds and quoted with its range (Appendix E). Of eight learned-router families, four re-run on the frozen matrix, trained and tuned on the calibration fold under CREDO's folds, labels and costs: the kNN router of Li (2025) and RouteLLM (Ong et al., 2024) (a text-router transplant) faithfully, AutoMix (Aggarwal et al., 2024) as a consistency-verifier variant, and CP-Router (Su et al., 2025) as a probe-score variant (its prediction-set rule needs option probabilities the matrix lacks); the other four (ECVL-Router (Tang et al., 2025), LatentRouter (Cheng & Dong, 2026), VL-RouterBench (Huang et al., 2025), RouterDC (Chen et al., 2024b)) need encoders or corpora the matrix cannot substitute for and enter as disclosed variants retrained on this pool's calibration fold (‡ rows; Appendix C). Random and oracle envelopes are exact matrix computations; only the CP-Router variant's kept branch carries a certificate. The main matrix cost 9.58 GPU-hours against a 64 GPU-hour cap (Appendix G).

4.2 RQ1: CERTIFIED COST-QUALITY FRONTIER

Figure 4 plots the cost-quality plane per benchmark and Table 1 pools it, adding the certificate axis. Certificate coverage varies widely by point: pooled over benchmarks it climbs with the risk level, from 17.2% of routed traffic at α=0.05 and 29.1% at α=0.10 to 58.1% at α=0.3; across the 30 (benchmark, α) operating points it runs 9.3%–75.8%, with 8 above one half. There is no cost headline: with the probe charged, no CREDO point undercuts the always-strongest tier. Its cheapest point, 3.53 s/item at α=0.3, is the probe's 3.07 plus routing, against 1.26 for the strongest tier; the largest saving versus always-strongest at matched-or-better pooled accuracy is -157.5% on the published split and lies between -160.0% and -157.5% across the 9 splits, and a point that both passes its own validity audit and costs less than the strongest tier exists on 0 of 9 splits. Nor is there an accuracy gain against the best fixed tier of each test fold, selected with test labels (an oracle comparator). Against the 32B tier, OCRBench at α=0.05 gains +4.8 to +8.0 points on every split; but the 32B is not the best fixed tier there; that is the 7B model on OCRBench, the 7B on HallusionBench, and the 32B is displaced on 20 of 54 (split, benchmark) folds, and against it the same point gains +0.4 points (-1.0 to +1.2 across splits, at least matching it on 7 of 9) at 6.7× its cost (6.5 to 6.8×), while clearing its own certified-risk budget on 2 of 9 splits (Appendix E). Over all 270 (benchmark, α, split) combinations routed accuracy exceeds the best fixed tier by at most +1.6 points (OCRBench at α=0.2) and trails it by as much as -6.8 (RealWorldQA at α=0.3); none reaches the 3-point floor we require for a claim, and pooled over benchmarks routing at α=0.05 lands -1.0 points below the best-fixed-tier policy (-2.0 to -0.5) at 3.84 against 1.01 s/item (Table 8).

The uncertified learned baselines are strong on raw cost-quality: the kNN router and the RouteLLM transplant land at 3.63–3.79 s/item at 67.2% pooled accuracy, inside CREDO's 3.53–3.84 s/item band but with no finite-sample statement, and the disclosed variants land in the same certificate-free band. The one certified baseline, the CP-Router probe-score variant, costs 4.02 s/item against CREDO's 3.75 on the published split, a gap near the top of the 9 for CREDO; across all of them CP-Router costs 4.00–4.05 and CREDO 3.75–3.98 s/item, cheaper on every paired split by 0.053–0.280 (median 0.158; a fold-sensitivity range, not an interval). The certificate axis favours CP-Router: its certified-branch risk is 0.145 (767) against CREDO's 0.242 here, and lower on every split (Appendix E), a trade-off on two axes rather than a ranking. COIN (Wang et al., 2025b), a calibrator that bounds the deployer's quantity directly, run through the same cascade at α=0.1, certifies 0.0–8.8% of traffic across the 9 splits against CREDO's 29.1%, at selective error 0.033–0.085 and 3.92–4.19 s/item, and certifies nothing in 988 of 1080 (benchmark, α, split, tier) thresholds. **Takeaway (RQ1): the budget raises coverage, not accuracy, and does not lower cost: certified share rises from 17.2% to 58.1% of traffic, no point undercuts the strongest tier once the probe is charged, and no point beats the best fixed tier of its own test fold by more than +1.6 points.**

4.3 RQ2: CERTIFICATE VALIDITY

The audit is two-sided. The *constructed* guarantee (Eq. 1) bounds each tier's false-accept rate on the questions it gets wrong; on test folds the realized per-tier rates straddle α, exceeding it at 57 of 120 (tier, benchmark, α) cells, with no vacuous calibration (Figure 1a). The bound is marginal, so the statistic it constrains is that pooled rate rather than a count of cells: over 4,443 wrong test answers it is 0.047, 0.104, 0.155, 0.201 and 0.292 at α=0.05, . . . , 0.3, at budget: a one-sided exact binomial test at level 0.05 places 0 of the five above α; over the 9 splits it places 3 of 45 (split, α) pairs

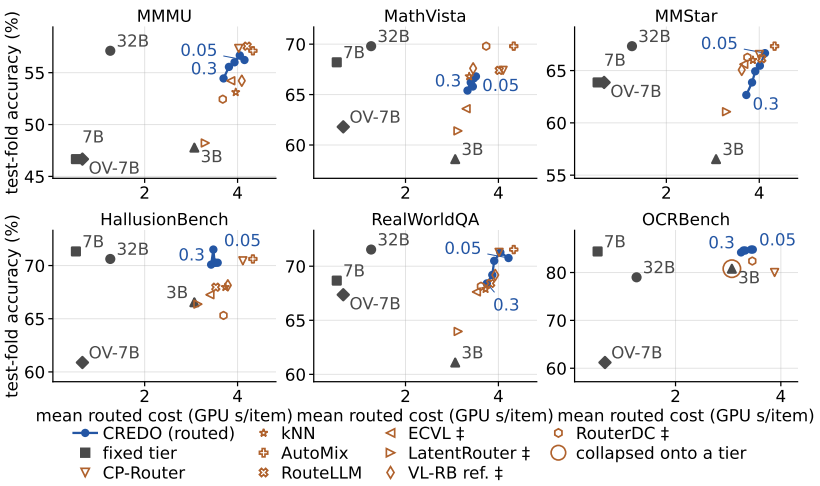

Figure 4: **RQ1: cost-quality operating points**, one panel per benchmark. Blue: CREDO's routed points (certified traffic pooled with strongest-tier fallback) in increasing α=0.05, . . . , 0.3, ends labelled; gray: fixed single tiers; open orange: the eight learned baselines (‡: disclosed variants, Appendix C). A ring around a tier marks learned routers that send every test query to that tier, so their point is the tier's: AutoMix on the 32B tier on all six benchmarks; RouteLLM, ECVL-Router and VL-RouterBench on the 7B tier on MathVista, HallusionBench and OCRBench; kNN on the 7B tier on HallusionBench and RealWorldQA. Same test fold throughout.

above, and each α's split mean is below it. The *deployment-facing* quantity, the conditional error of the traffic that ships with a certificate, is not bounded by that construction, and Table 2 shows the consequence: certified-traffic risk exceeds α at 23 of 30 operating points, split-robustly 23–25 across the 9 splits (Figure 7, Appendix E). The gap is not a labeling artifact: a three-stage audited wash of the label pipeline (Appendix A) left the violation pattern essentially unchanged, 24 of 30 before and 23 of 30 after. The count is a point estimate: some cells' intervals straddle their budget, and cells whose base rate puts the budget out of any probe's reach are counted with the rest. A pre-registered fifth tier from a third model family (Pixtral-12B, a Mistral-Nemo backbone; Appendix K) leaves the count within two operating points (23 of 30, +0 against the recomputed baseline), and the new tier's per-tier bound straddles α as the published four's do (14 of 30 cells; Table 15). **Takeaway (RQ2): the bound holds for the quantity it states (57 of 120 tier cells above α, straddling) and fails for the error a deployer observes (23 of 30 points).**

4.4 RQ3: PROBE ABLATION

We rescan the entire pipeline (scores, thresholds, routing, frontier) with each of the four channels (question features, the 3B draft's logprob/entropy pair, the k=5 vote entropy, CLIP alignment) removed or used alone, on identical folds and labels; discrimination is the AUROC of the probe score against correctness per (benchmark, tier) test-fold cell, pooled by test-fold size. Two findings (Figure 5; table in Appendix F). First, discrimination is weak everywhere: the full probe pools to an AUROC of 0.563 on the published split, already the sweep's best value, and even on the probe tier's own cells per-cell AUROC spans only 0.553–0.763. Second, in this pooling no single channel is necessary, and we decline to rank them: removing any single channel moves the pooled AUROC by at most 0.0141, to no lower than 0.548, less than the split-to-split scatter of the full probe (0.554–0.572 across 9 splits), so the leave-one-out ordering is not separable from resampling noise. Run alone the channels span a narrow band (Appendix F) in which two contrasts separate: logprob/entropy leads the question features, and CLIP alignment, the cheapest multimodal signal, is chance-level alone, its removal leaving the probe at 0.560. **Takeaway (RQ3): pooled over all cells no single channel is necessary: leave-one-out moves pooled AUROC by at most 0.0141, less than the 0.554–0.572 split scatter, and CLIP alignment alone is at chance (0.502); Section 5 disaggregates it.**

Table 1: **RQ1 summary: certified frontier vs. baselines**, pooled over the six test folds (per benchmark: Appendix J; cross-split: Appendix E). CREDO rows pool certified traffic with strongest-tier fallback; "Cert. risk" is the conditional error of certified traffic (not bounded by Eq. 1), "Meets target" the benchmarks with risk $\leq \alpha$ (point estimates); † marks re-runs adapted to the frozen matrix with a disclosed substitution (Appendix C), ‡ variants retrained on this pool's calibration fold; envelopes carry no certificate.

Method	Cert.	Mean cost (s/item)	Quality (%)	Cert. share (%)	Cert. risk	Meets target
CREDO α=0.05	✓	3.84	69.2	17.2	0.220	0/6
CREDO α=0.10	✓	3.75	69.0	29.1	0.242	1/6
CREDO α=0.20	✓	3.62	68.4	47.3	0.261	1/6
CREDO α=0.30	✓	3.53	67.4	58.1	0.286	4/6
CP-Router (probe-score variant)†, α=0.10	✓	4.02	68.8	24.4	0.145 (767)	1/6
kNN router	–	3.63	67.2	–	–	–
AutoMix†	–	4.13	69.5	–	–	–
RouteLLM transplant	–	3.79	68.0	–	–	–
ECVL-Router (variant)‡	–	3.48	66.6	–	–	–
LatentRouter (variant)‡	–	3.19	63.7	–	–	–
VL-RouterBench ref. (variant)‡	–	3.64	67.5	–	–	–
RouterDC (variant)‡	–	3.67	67.5	–	–	–
Random envelope	–	1.38	64.7	–	–	–
Oracle envelope	oracle	0.64	86.1	–	–	–

Table 2: **RQ2: validity audit.** Conditional error of certified traffic versus budget α, per benchmark, with the certified-subset size in parentheses; **bold** marks a violation (risk $> \alpha$), reported, never patched. Cells are point estimates on the published split; read them with the split-robust count (23–25/30, Appendix E).

Benchmark	α=0.05	α=0.10	α=0.15	α=0.20	α=0.30
MMMU	**0.288** (73)	**0.347** (118)	**0.364** (162)	**0.388** (209)	**0.413** (254)
MathVista	**0.271** (96)	**0.232** (177)	**0.239** (226)	**0.243** (268)	0.277 (325)
MMStar	**0.252** (143)	**0.323** (217)	**0.327** (281)	**0.344** (320)	**0.375** (400)
HallusionBench	**0.321** (53)	**0.296** (115)	**0.305** (187)	**0.279** (215)	0.295 (278)
RealWorldQA	**0.200** (40)	**0.172** (99)	**0.201** (144)	**0.237** (156)	0.244 (193)
OCRBench	**0.081** (135)	0.099 (191)	0.108 (287)	0.112 (322)	0.129 (379)

4.5 RQ4: POOL SCALABILITY (SINGLE SPLIT)

We remove one tier at a time and re-trace the frontier from the cached matrix on the same folds, on the published split only (Table 3). The probe is paid whether or not tier 1 stays in the pool, so dropping tier 1 forfeits its free answers and *raises* pooled cost, by +4.5% shipped and +6.2% clean, while weakening the certificate: the logprob/entropy and vote channels come from that forward pass, so abstention rises from 70.9% to 86.9% and the reduced pool certifies less traffic. Dropping the strongest tier lowers pooled cost most under either pricing (-7.2% shipped, -9.9% clean). **Takeaway (RQ4, single split): pool composition moves where uncertified traffic lands (abstention 70.9% to 86.9% without tier 1), not whether the certificate holds.**

5 ANALYSIS

The gap is a base-rate identity, and the calibration fold predicts it. For one tier, the quantity Eq. 1 bounds and the quantity a deployer reads are linked exactly. With FA the tier's false-accept rate $P(\text{accept} \mid \text{wrong})$, TA its true-accept rate $P(\text{accept} \mid \text{correct})$, and π its marginal wrong rate,

$$P(\text{wrong} \mid \text{accept}) \;=\; \frac{\text{FA}\,\pi}{\text{FA}\,\pi + \text{TA}\,(1-\pi)}. \tag{3}$$

Calibration caps FA at α; the certified error is then decided by TA against π. A probe at AUROC one half accepts wrong and correct answers at the same rate, TA = FA, and Eq. 3 returns π whatever

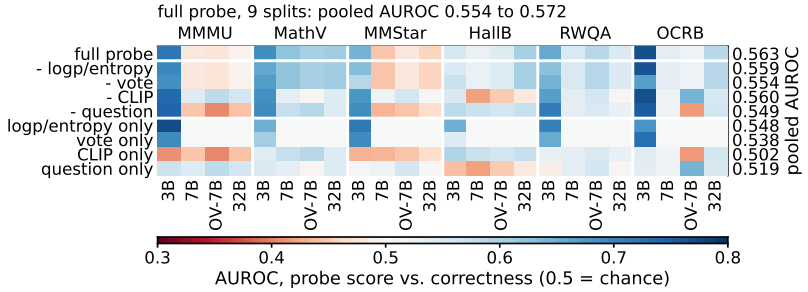

Figure 5: **RQ3: probe-channel rescans.** AUROC of the probe score against correctness per (benchmark, tier) test-fold cell (columns) for the full probe, four leave-one-out and four single-channel variants (rows), diverging around chance; right margin, each variant's pooled AUROC, whose 9-split span 0.554–0.572 is wider than any leave-one-out difference.

Table 3: **RQ4: pool scalability.** Certified frontier with one tier removed, re-traced on identical folds, published split only. Costs use the shipped tier-1 constant (clean-constant repricing shifts magnitudes only); abstain rate is the abstain-policy counterpart's.

Configuration	Mean cost @ α=0.10 (s/item)	Abstain rate (%)
Full pool (4 tiers)	3.75	70.9
− tier 1 (3B)	3.92	86.9
− tier 2 (7B)	3.90	71.1
− tier 3 (OV-7B)	3.76	71.5
− tier 4 (32B)	3.48	70.9

α was; the certified error falls below π only as fast as the probe raises TA above α. HallusionBench at α=0.2 shows it in one row: all four tiers' test-fold false-accept rates lie in 0.150–0.170, under budget, and the certified error is 0.279. Every term of Eq. 3 is available on the calibration fold, so the gap can be predicted before any test label is read: we replay the cascade there with the thresholds fit on it and take its certified error as the prediction. Figure 6a compares prediction and measurement at the 30 operating points of the published split: Pearson r=0.90 (0.83–0.95 across the 9 splits), mean absolute deviation 0.039; the calibration fold's verdict (certified error above or below α) matches the test fold's at 27 of 30 points (26–30 across splits), predicting 24 violations against 23 measured. Both statistics need a floor: a constant "always violation" matches at 23 of 30, and the probe-free rule $\bar{\pi}_{\text{cal}} \geq 0.3$ at 26 with r=0.85, so the replay's increment over them is small, and what the identity provides is the mechanism, not a sharper predictor. Per (tier, benchmark, α) cell the same prediction gives r=0.69 over 120 cells, mean absolute deviation 0.069 (Table 9).

The probe sets where the error sits between α and π. Eq. 3 also says where the certified error sits between the budget α (a perfect probe) and the chance limit π (an uninformative one). Per cell whose chance limit lies at least one grid step above the budget, the gap fraction $g = (e-\alpha)/(\pi-\alpha)$ places the measured accepted-region error e on that scale. The identity predicts g falls as the probe separates correct from wrong answers; Figure 6b measures it against each cell's AUROC over 24 cells: Spearman $\rho = -0.94$, median g=0.68. The sign is implied by Eq. 3 (false accepts pinned near α, g falls as true accepts rise); the magnitude and the spread of g, on one test fold, are the empirical content. The least separable cell, MMStar/7B at AUROC 0.442, sits above chance (g=1.69: the probe ranks wrong answers higher); the most separable, OCRBench/3B at 0.763, is below budget ($g = -0.42$), and OCRBench is the one benchmark whose certified error meets its budget at $\alpha \geq 0.1$ (Table 2). The oracle envelope (each question to the cheapest tier that answers it) reaches 86.1% at 0.64 s/item against 70.2% for the best fixed tier per benchmark (Table 1), and of those 15.9 points the probe, at AUROC 0.563, recovers none.

A better probe narrows the gap and does not close it. Pre-registered extension B4 (Appendix L) replaces the rule probe's score with a logistic probe on the 3B tier's last-layer hidden state from one prefill, trained on half of the calibration fold. Last-token pooling clears the registered signal bar (pooled AUROC on the published split 0.718 against the rule probe's 0.563; 4 of six benchmarks at

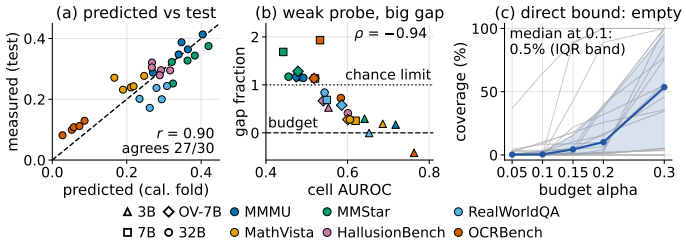

Figure 6: **The base-rate mechanism, predicted and then measured.** (a) Certified-traffic error predicted from the calibration fold (the cascade replayed on the fold its thresholds were fit on) against the test-fold measurement, one marker per (benchmark, α). (b) Per (benchmark, tier) cell, the probe's AUROC against the mean gap fraction g (defined below). (c) Test-set-optimal coverage of an accepted-region (RCPS-style) certificate per cell. Markers are tiers, colours benchmarks.

or above 0.70 on the 7B tier); mean pooling does not (0.629, 1 of six). On the same 30 operating points the count over budget moves from 23 to 20 (20–25 across the nine splits) with last pooling and to 26 with mean pooling, against a registered target of at most 15 and a registered reading of "not probe-limited" at 20 or more. The rule fires on the second branch. At its scope that says these two probes at this budget do not remove over-admission rather than that no probe could: a pooled AUROC does not fix the tail ratio a threshold selects. The matched comparison is the halved-fold rule probe's 24 against last pooling's 20.

The pooled ablation dilutes its own channels. Read off the probe tier's forward pass, the log-prob/entropy and vote channels exist on only six of the cells Section 4.4 pools. Re-pooled there the full probe reaches 0.664, dropping the vote channel costs 0.036 (wider than its split scatter), the question rule is subtractive (0.688 without it, 0.692 for logprob/entropy alone, the sweep's best), and a 7B probe tier leaves the pooled AUROC at 0.551 against 3B's 0.560 on the same rows (Appendix H). The RQ3 null is about the pooling, not the channels: a per-benchmark logistic fusion of the same inputs, fit on half of the calibration fold, lifts pooled test AUROC, averaged over 9 splits, from 0.561 to 0.679 (permuted labels 0.494), and the certificate built on it is still over budget at 21–23 of 30 operating points (Appendix M).

Not the construction. Certifying the accepted region directly bounds the deployment-facing error by definition; Figure 6c shows its coverage: at α=0.10 the median cell certifies 0.5% of its traffic, 15 of 24 cells at most 1% (Appendix D).

Negative results. Absent and reported rather than patched (RQ1–RQ3): a cost saving at matched accuracy once the probe is charged, an accuracy gain over a test fold's best fixed tier, a separable channel ranking, a label-noise explanation. What remains is the base rate against a weak probe (B4).

Scope and limitations. The bound in Eq. 1 is marginal over the test distribution and conditional on exchangeability of the two folds; it is not a per-query guarantee, and drift requires recalibration. It covers answer-quality risk only; the cost model is GPU-seconds on one serving stack; the benchmarks are English VQA and the nine splits resample one frozen matrix. On the design-external AI2D (Appendix I) the thresholds hold at 4 of 5 α points, coverage +18.7 pp above the design mean.

6 Conclusion

CREDO brings finite-sample, distribution-free certificates to routing over VLM pools. Routing raises coverage, not accuracy, and does not lower cost: it never gains more than +1.6 points on the best fixed tier, and with the probe charged no point undercuts the strongest tier. The error of certified traffic exceeds its budget at 23 of 30 operating points. A per-tier false-accept bound does not bound the error of the traffic it admits. Weak-signal channels pool to at most 0.572 AUROC on any split under the shipped rule, a learned fusion of them reaches 0.679 and leaves 21–23 of 30 points over budget, and a hidden-state probe above the registered bar leaves 20 of 30 against the rule probe's 24 on the same halved fold (Appendices H, L, M).

ETHICS STATEMENT

All data and models used here are public, pinned, and used within their licences; no human subjects or private data are involved. The main ethical risk the method itself carries is misreading: a certificate bounds a marginal, per-tier false-accept rate, not the correctness of any single delivered answer, and the audit in Section 4.3 shows that the conditional error of certified traffic can exceed the budget the user set. We state the marginal semantics wherever the bound appears and report every violation rather than patching it, so that a deployer cannot mistake the guarantee for more than it is.

REPRODUCIBILITY STATEMENT

All benchmarks, models, and revisions used in this paper are pinned to literal hub API responses (Appendix A); the full response matrix, per-tier prompts, official-scorer ports, calibration splits, and the main-matrix GPU ledger are in the supplementary archive (`README_supp.md` at its root; the matrix under `experiments/credo/results/raw/`). Every result numeral in this paper is generated from the released result artifacts by a macro bridge (`gen_macros.py` reading `frontier.json`, `ablation.json`, `baselines.json`, and the pre-audit frontier), with no hand-transcribed numbers; the learned-baseline re-runs are calibration-fold-only CPU rescans of the same frozen matrix (Appendix C), and the four baseline families whose faithful constructions cannot be reproduced from it enter the baseline table only as transparently disclosed calibration-fold variants, never under their original names alone.

AI USE STATEMENT

In this work we disclose generative AI use under the categories of the ICLR 2027 AI policy. Within the required-disclosure categories, we used generative AI tools to help develop the conceptual framework, propose and refine hypotheses, formulate the paper's formal claims and statistical analyses, give feedback on research methodology and experiment design, implement methods (the experiment, analysis and figure-generating code released with this work), clean and reformat datasets, and interpret results. We have *not* used generative AI tools to generate synthetic data sets, assist with translation, support qualitative or thematic data analysis, or write proofs (this paper states no formal proof); the remaining required-disclosure categories are not applicable to this work, which has no human-subject component and no data collection beyond public benchmarks and the outputs of open-weight models. Under the recommended categories, generative AI was additionally used for software code, literature search and analysis, brainstorming, figure preparation, and editing the text of this paper. No experimental result was produced, selected or filtered by a generative AI tool: every experimental numeral in this paper is emitted by a committed single-writer artifact and reaches the text through an auto-generated macro bridge that admits no hand-typed value; design constants, protocol settings and figures quoted from prior work are of course written by hand and cited. The research questions, the claims and the conclusions are the authors'. We have reviewed all AI-assisted work and take responsibility for the final content, including every claim, number and artifact.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Ankit Anand, et al. Automix: Automatically mixing language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511, 2021. Foundations and Trends in Machine Learning.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Anonymous. Conformal risk-controlled routing for large language model. OpenReview submission lLR61sHcS5, 2025. under review; https://openreview.net/forum?id=lLR61sHcS5.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report. arXiv:2502.13923, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, et al. Are we on the right way for evaluating large vision-language models? arXiv:2403.20330, 2024a. NeurIPS 2024 D&B per task brief.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. arXiv:2305.05176, 2023. later TMLR 2024.

Shuhao Chen, Weisen Jiang, Baijiong Lin, James T. Kwok, and Yu Zhang. Routerdc: Query-based router by dual contrastive learning for assembling large language models. arXiv:2409.19886, 2024b. NeurIPS 2024 per task brief.

Xueqi Cheng and Yushun Dong. Latentrouter: Can we choose the right multimodal model before seeing its answer? arXiv:2605.11301, 2026.

Dujian Ding, Ankur Mallick, Chi Wang, et al. Hybrid llm: Cost-efficient and quality-aware query routing. In *International Conference on Learning Representations (ICLR)*, 2024.

Yifan Dou, Shikan Lian, and Shibo Li. Conformal cascade: Distribution-free accuracy guarantees for multi-tier llm inference. arXiv:2607.25018, 2026.

Ayush Dwivedi, Qixin Wang, Ashvi Soni, et al. Safe-cascade: Cost-adaptive vision-language routing for chart question answering. arXiv:2606.19646, 2026.

Arya Fayyazi and Haleh Akrami. Proof-of-perception: Certified tool-using multimodal reasoning with compositional conformal guarantees. arXiv:2603.00324, 2026.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. arXiv:2310.14566, 2024. CVPR 2024 per task brief.

Dongxin Guo, Jikun Wu, and Siu Ming Yiu. Routenlp: Closed-loop llm routing with conformal cascading and distillation co-optimization. arXiv:2604.23577, 2026.

Sai Hao, Hao Zeng, Hongxin Wei, and Bingyi Jing. Racer: Risk-aware calibrated efficient routing for large language models. arXiv:2603.06616, 2026.

Qiushuo Hou, Sangwoo Park, Matteo Zecchin, et al. Reliable llm-based edge-cloud-expert cascades for telecom knowledge systems. arXiv:2512.20012, 2025.

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:2403.12031, 2024.

Zhehao Huang, Baijiong Lin, Jingyuan Zhang, et al. Vl-routerbench: A benchmark for vision-language model routing. arXiv:2512.23562, 2025. arXiv:2512.23562; task brief lists CVPR 2026, venue unconfirmed.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: Llm judges with provable guarantees for human agreement. In *International Conference on Learning Representations (ICLR)*, 2024. ICLR 2025.

Varun Kotte. Ucci: Calibrated uncertainty for cost-optimal llm cascade routing. arXiv:2605.18796, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, et al. Llava-onevision: Easy visual task transfer. arXiv:2408.03326, 2024.

Yang Li. Rethinking predictive modeling for llm routing: When simple knn beats complex learned routers. arXiv:2505.12601, 2025.

Yuliang Liu, Zhang Li, Mingxin Yin, et al. Ocrbench: On the hidden mystery of ocr in large multimodal models. arXiv:2305.07895, 2024.

Zhongye Liu, Yaopei Zeng, Yurui Chang, and Lu Lin. Forced deferral: Manipulating routing decisions in multimodal llm cascades. arXiv:2606.15308, 2026.

Pan Lu, Hritik Bansal, Tony Xia, et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv:2310.02255, 2024. ICLR 2024 per task brief.

Haoxuan Ma, Guannan Lai, and Han-Jia Ye. Mmr-bench: A comprehensive benchmark for multimodal llm routing. arXiv:2601.17814, 2026.

Yasmin Moslem and John D. Kelleher. Dynamic model routing and cascading for efficient llm inference: A survey. arXiv:2603.04445, 2026.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. arXiv:2406.18665, 2024.

William Overman and Mohsen Bayati. Conformal arbitrage: Risk-controlled balancing of competing objectives in language models. arXiv:2506.00911, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. arXiv:2103.00020, 2021. ICML 2021 per task brief.

Ahmed Šabanović, Paul Joe Maliakel, and Ivona Brandić. Inar-vl: Input-aware routing for edge-cloud vision-language inference. arXiv:2605.18853, 2026.

Jiayuan Su, Fulin Lin, Zhaopeng Feng, et al. Cp-router: An uncertainty-aware router between llm and lrm. arXiv:2505.19970, 2025. AAAI 2026 per task brief.

Xin Tang, Youfang Han, Fangfei Gou, et al. Ecvl-router: Scenario-aware routing for vision-language models. arXiv:2510.27256, 2025.

Mikael Turkoglu, Tim Bary, Vincent Thielens, Manon Dausort, and Benoît Macq. Cads: Conformal adaptive decision system for cost-efficient image classification. arXiv:2605.16401, 2026.

Antonios Valkanas, Soumyasundar Pal, Pavel Rumiantsev, Yingxue Zhang, and Mark Coates. C3po: Optimized large language model cascades with probabilistic cost constraints for reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zhiyuan Wang, Aniri, Tianlong Chen, et al. Lec: Linear expectation constraints for selection-conditioned risk control in selective prediction and routing systems. arXiv:2512.01556, 2025a.

Zhiyuan Wang, Jinhao Duan, Qingni Wang, Xiaofeng Zhu, Tianlong Chen, Xiaoshuang Shi, and Kaidi Xu. COIN: Uncertainty-guarding selective question answering for foundation models with provable risk guarantees. arXiv:2506.20178, 2025b.

xAI. Grok-1.5 vision preview. `https://x.ai/news/grok-1.5v`, 2024. introduces the RealWorldQA benchmark; dataset mirrored at huggingface.co/datasets/xai-org/RealworldQA.

Yifan Xiao, Zhongjun Wang, Tiantian Yan, Weiwei Li, Dazhong Wu, and Jie Ma. Time series anomaly detection: A hybrid deep learning framework with adaptive thresholding for electricity industry data. In *IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)*, 2025. DOI 10.1109/CSCLOUD66326.2025.00046.

Yifan Xiao, Shijie Li, and Yuhao Ge. RunbookFX: Type- and effect-safe LLM synthesis for executable incident diagnosis and mitigation. *Proceedings of the ACM on Programming Languages*, 10(ICFP), 2026a. Article 277; DOI 10.1145/3828675.

Yifan Xiao, Shijie Li, and Yu Huang. CALM-TS: Risk-controlled LLM labeling for time-series via calibrated selective gating. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026b. DOI 10.1145/3770855.3817859.

Dexuan Xu, Yanyuan Chen, Zhongyan Chai, Yifan Xiao, Yandong Yan, Weiping Ding, Hanpin Wang, Zhi Jin, Wenpin Jiao, Weihua Yue, Hang Li, and Yu Huang. Knowledge fusion in deep learning-based medical vision-language models: A review. *Information Fusion*, 125:103455, 2026. doi: 10.1016/j.inffus.2025.103455.

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. Large language model cascades with mixture of thoughts representations for cost-efficient reasoning. In *International Conference on Learning Representations (ICLR)*, 2024a.

Xiang Yue, Yuansheng Ni, Kai Zhang, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv:2311.16502, 2024b. CVPR 2024 per task brief.

A REPRODUCIBILITY DETAILS

Pinned assets. Every external asset is pinned to the literal revision returned by the hub API on 2026-07-18 (full evidence in the released `runbook/DATA_SOURCES.md`). Benchmarks: MMMU (`MMMU/MMMU`, 30 subject configs, validation), MathVista (`AI4Math/MathVista`, testmini), MMStar (`Lin-Chen/MMStar`, val), HallusionBench (`lmms-lab/HallusionBench`, image+non_image), RealWorldQA (`xai-org/RealworldQA`, test), OCRBench (`echo840/OCRBench`, test). Pool: Qwen2.5-VL-3B/7B/32B-AWQ and LLaVA-OneVision-7B at pinned snapshots; the offline loader hard-stops on any pin-versus-disk mismatch. Calibration split: seeded permutation, seed 20260718, 50/50. Table 4 gives the pool in cascade order with its smoke-measured per-item costs and Table 5 the six benchmarks with their official splits and sizes.

Table 4: The CREDO pool: four frozen VLMs in cascade order. Cost is the smoke-measured GPU-seconds per item on the serving stack; tier 1's cost includes its probe forward, which inverts the design-time cost ranking, and its constant is conservative (Appendix E).

Tier	Model	dtype	Cost (s/item)
1 (probe)	Qwen2.5-VL-3B-Instruct (Bai et al., 2025)	bf16	3.07
2	Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	bf16	0.52
3	LLaVA-OneVision-7B (Li et al., 2024)	bf16	0.66
4 (strongest)	Qwen2.5-VL-32B-AWQ (Bai et al., 2025)	fp16	1.26

Table 5: Benchmarks, official splits, and sizes (pinned hub revisions, verified 2026-07-18). Scoring is the benchmark's official logic ported 1:1; HallusionBench's GPT judge step is replaced by a deterministic yes/no judge, and RealWorldQA has no official scorer so the de-facto exact-match protocol is used (Appendix B).

Benchmark	Split	#Q	Answer format
MMMU (Yue et al., 2024b)	validation	900	MC / open
MathVista (Lu et al., 2024)	testmini	1,000	MC / numeric
MMStar (Chen et al., 2024a)	val	1,500	MC
HallusionBench (Guan et al., 2024)	image+non-image	1,129	yes/no
RealWorldQA (xAI, 2024)	test	765	MC / short
OCRBench (Liu et al., 2024)	test	1,000	short text
Total		6,294	

Prompts and generation. Each benchmark is prompted in its official format (e.g., MathVista's released `query` field is used verbatim; OCRBench generates at most 100 new tokens per its official example). All tiers decode greedily; the probe tier additionally emits top-20 logprobs and $k{=}5$ self-consistency samples at temperature 0.7. The generation cap is 512 new tokens on MathVista (all tiers) and on the MMMU cell of the 7B tier, and 128 elsewhere: the initial 128-token pass truncated most of the CoT-style tiers' MathVista outputs before their final answer, 59–88% of them, a diagnostic measured on that pass and recorded in the run log rather than in a released artifact, since those outputs were replaced by the regeneration it prompted, and a scoring-integrity audit caught the resulting per-tier accuracy inversion, so only the affected cells were regenerated at 512 and every cell was rescored under one fixed scoring protocol. The released `scoring_audit.json` carries the accuracies of both passes per tier, which is what the argument rests on; the truncation fractions are the reason the regeneration happened, not a result the released matrix reproduces. This is the three-stage label wash referenced in Section 4.3: (1) scorer-side fix of the inversion, (2) truncation-repair regeneration, (3) full rescore. The audit artifact with per-cell truncation rates before and after ships with the results (`results/scoring_audit.json`), and the pre-wash frontier is released unmodified as `results/frontier_v2_pre-audit.json` so the persistence of the RQ2 violations (24/30 before vs. 23/30 after) is independently checkable.

Tier-4 sanity gate. The AWQ-quantized 32B tier is screened by a mandatory in-code gate: the first 20 outputs of every tier-4 cell are checked for overflow garbage (threshold 5%), and a tripped

gate aborts to a documented bf16 fallback on a solo card; the script never generates past a tripped gate.

Ledger. GPU hours are guarded before every main-matrix model load and recorded in a finally block, so burned hours are real even on crashes; for those runs the cap of 64 GPU-hours is enforced in code. The ledger covers the main matrix, whose 9.58 GPU-hours Section 4.1 reports. The extensions of Appendices K, H, L and I and the cost-constant recheck ran outside it and are not in that total.

Baseline nondeterminism, and what it does to Table 1. The learned-baseline pass is not a deterministic function of the frozen matrix, so we measure the effect and ship the measurement rather than assume it away. Every tunable family picks its operating point by scoring a knob grid on calibration-fold accuracy and cost, and those scores carry exact ties (z-scored small-integer features produce exactly tied neighbour distances and tied fold accuracies), which float reductions break by intermediate-buffer alignment: deterministic for a fresh interpreter's allocation sequence, not across repeated in-process calls. On the recording host, three repeat draws of the pass inside one interpreter return different pooled costs for seven of the nine families, the widest being 0.015 s/item or 0.5% of that family's published value. The mechanism is a discrete knob flip on MMMU (a sufficiency threshold moving from 0.65 to 0.50, a utility weight from 1.2 to 2.0) rather than a last-place rounding wobble, and the draws do not settle to a warm value. **The learned-baseline rows of Table 1, and every cross-split baseline cell derived from them, are therefore one draw from this procedure, and repeat draws of that procedure disagree at the percent level.** CP-Router and AutoMix are exactly stable across the repeat draws, which is why the certificate comparison the paper leans on (Section 4.2, Appendix E) is not exposed to this. Two mitigations are engineered rather than argued: each per-seed pass in Appendix E runs in a fresh subprocess, which reproduces the committed `baselines.json` bit-for-bit here; and the released no-drift assertion compares floats to a tolerance and discrete knobs exactly, with a failure message that names this nondeterminism, so a reproducer on other hardware reads a flipped knob as the documented behaviour it is, not as a corrupted artifact. The probe regenerates with the artifact (`baseline_nondeterminism` in `results/split_sensitivity.json`).

B SCORING PORTS

Per-item labels come from each benchmark's official evaluation logic, ported 1:1 and unit-tested on hand-computed fixtures: MMMU's `eval_utils` answer extraction; MathVista's normalization, choice matching, and deterministic extraction, with the official pipeline's GPT-4 extraction fallback replaced by a deterministic approximation (last answer-indicator phrase or `\boxed{}` payload, last option-letter or verbatim-choice mention, last number) that engages only where the official deterministic rules fail, never consults the gold answer, and applies identically to every model; a neutrality control (the direct-answer 3B tier moves under the same rules) and a 30-item human validation are in the released audit artifact; MMStar's prefix rule; OCRBench's substring and HME100k special cases with the official /1000 final-score semantics; HallusionBench's aAcc/qAcc/fAcc chain with the same GPT-judge step replaced by a deterministic yes/no judge over the benchmark's binary answer space (documented substitution). RealWorldQA ships no official scorer; we implement the de-facto VLMEvalKit/lmms-eval exact-match protocol and label it as such.

C IN-POOL LEARNED-BASELINE IMPLEMENTATIONS

All learned baselines are CPU-only re-runs on the frozen response matrix and manifests, with no new generations, released as `results/baselines.json` (single writer: `baselines.py`; per-family implementation notes and deviations ship in its `_meta` fields). Common protocol: the identical seeded 50/50 calibration/test split as `frontier.json` (fold sizes asserted in code against it); training and threshold tuning touch calibration-fold labels only, test labels appear exclusively in the fenced evaluation block; each routed query is charged the probe forward (3.07 s/item) plus the smoke-measured sec/item of any higher tier that answers it (the RQ1 cost convention; tier 1's answer is the draft the probe already paid for). Features are the signals the frozen artifacts carry: the twelve deterministic question features, the CLIP alignment score (HallusionBench's imageless questions carry none; the baselines impute the calibration-fold mean, with the image-metadata flag

as indicator), and the 3B probe channel (mean logprob, mean entropy, vote entropy). Tunable baselines pick their operating point on the calibration fold by a uniform rule: the cheapest threshold whose calibration-fold routed accuracy is at least the strongest tier's, else the most accurate one, a label-informed tuning step CREDO itself never gets.

kNN router (Li, 2025): per-tier correctness estimated as the mean over the k=32 nearest calibration questions in z-scored feature space (leave-one-out on the calibration fold); route to $\arg\max_m[\hat{p}_m(x) - \lambda\, c_m]$ with λ tuned as above. **RouteLLM transplant** (Ong et al., 2024): pairwise cheapest-versus-strongest; logistic regression on the features predicts the cheap tier's correctness, routing down when the predicted probability clears the tuned threshold (the original's preference data and BERT/matrix-factorization scorers are replaced by correctness labels and matrix features). A gradient-boosted variant ships in the artifact under `routellm_gbm`; where a family offers more than one scorer, Table 1 shows the higher-accuracy one, which here is the logistic variant; the gradient-boosted one is cheaper and weaker (3.20 s/item at 63.7% pooled accuracy, versus 3.79 at 68.0%). That selection is not an artifact of the published fold: under the 9 splits of Appendix E the logistic variant is the more accurate of the two on every split (Table 7, which carries both rows). **AutoMix** (Aggarwal et al., 2024), consistency-verifier variant†: the 3B answers and escalates to the strongest tier when its k=5 self-consistency agreement falls below the tuned threshold; the original's few-shot self-verification prompt would require new generations, so the matrix's self-consistency channel stands in and the POMDP meta-verifier reduces to a threshold. Under the measured rates the 3B is dearer per item than the 32B, so the tuned cascade collapses to always-strongest, reported as measured. **CP-Router** (Su et al., 2025), pairwise probe-signal variant†: the same wrong-subset split-conformal construction as CREDO, restricted to the (3B, strongest) pair at α=0.10 and driven by the 3B's own uncertainty channels only; the original thresholds MCQ prediction-set size from option-level probabilities, which the matrix does not store, so the probe-channel score stands in. Its kept branch is the one baseline operating point with a finite-sample false-accept certificate.

Faithful constructions not reproducible; disclosed variants instead (‡ rows in Table 1): four families need a trained encoder, latent representation, or training corpus that no signal in the matrix or manifests can substitute for: ECVL-Router trains a response-quality predictor on its RSD image-text corpus over scenario objectives; LatentRouter learns multimodal routing capsules from raw image-question inputs; the VL-RouterBench reference routers are trained on that benchmark's own sample-model log corpus over a seventeen-model pool; RouterDC trains an mDeBERTa query encoder with dual contrastive losses. The artifact records these reasons verbatim (`faithful_not_reproducible`) alongside the exact additional passes (e.g. a query-embedding pass for RouterDC, a self-verification pass for faithful AutoMix, option-level scoring for faithful CP-Router) that would unblock a faithful run. What enters Table 1 instead are calibration-fold-retrained counterparts over the same frozen features, reported only under disclosed variant names, never as the original methods' numbers. **ECVL-Router** (Tang et al., 2025), feature-space variant‡: per-tier logistic quality predictors fit on the calibration fold stand in for the corpus-trained response-quality predictor; the scenario objective reduces to a sufficiency rule: route to the cheapest tier whose predicted quality clears the tuned threshold τ, else the predicted-best tier. **LatentRouter** (Cheng & Dong, 2026), feature-space variant‡: a shared-trunk MLP regressor (seeded) predicts the per-tier correctness vector jointly (the shared hidden layer standing in for the latent routing representation, the per-tier output heads for the model-capability tokens) and routes to $\arg\max_m[\hat{q}_m(x) - \lambda\, c_m]$ with λ tuned as above. Disclosed limitation: the trunk is fitted by L-BFGS under a 500-iteration cap and hits that cap on all six benchmarks, so this row is an *unconverged* fit, reported as run rather than tuned until it converges (raising the cap is CPU-cheap but would move a published number, so it is disclosed here instead). It is the lowest-accuracy family in Table 1, and its cross-split accuracy is in Table 7; its position in the certificate-free band is therefore weak evidence about the construction it stands in for. **VL-RouterBench reference protocol** (Huang et al., 2025), pool-retrained variant‡: the benchmark's predict-quality-then-route-under-a-cost-cap protocol with its released encoders and log corpus replaced by per-tier logistic predictors retrained on this pool's calibration fold; route to the predicted-best tier among those under the tuned per-query cost cap. **RouterDC** (Chen et al., 2024b), feature-space variant‡: a linear query encoder and learned per-tier embeddings trained with the original's sample-LLM and sample-sample contrastive losses (seeded k-means clusters, full-batch gradient descent) on the calibration fold; route to the top-scoring tier with no cost term, as in the original.

D WHY NOT CERTIFY THE ACCEPTED REGION DIRECTLY?

Eq. 1 bounds the false-accept rate *given wrong*; the deployment-facing quantity RQ2 audits, the error rate *inside* the accepted region, could instead be controlled directly by a risk-controlling prediction-set / conformal-risk-control construction (Angelopoulos et al., 2024) that calibrates the acceptance threshold against the accepted region's conditional error. We did not switch constructions, and the released oracle sweep (`results/ablation.json`, `rcps_oracle`) is the reason. Per (benchmark, tier) cell we order the test fold by the full probe score and find the largest acceptance region $\{s > t\}$ whose realized conditional error is $\leq \alpha$, an upper bound on the coverage of any region that meets α on this test fold, since it picks the threshold with test labels in hand. At α=0.05, 18 of 24 cells can certify at most 1% of their traffic; at α=0.10 it is 15 of 24 cells with median certifiable coverage 0.5% (the best cell, OCRBench on the probe tier, reaches 64.6%); at α=0.15 the median is still 4.4%. A certificate that ships almost no traffic is not a usable routing primitive, so the false-accept construction plus an honest audit of its deployment-facing gap (Section 4.3) is the deployable choice, and the collapse itself corroborates RQ3: with probe scores this weakly informative, *no* threshold construction can carve out a large low-error region.

The sweep also quantifies the tie pathology discussed in Section 5: probe scores tie heavily on multiple-choice suites, acceptance regions $\{s > t\}$ can only grow in tie-block jumps, and oracle-certifiable coverage accordingly moves discontinuously: on one (benchmark, tier) cell it leaps from 2.3% to 100.0% between the adjacent grid budgets α=0.2 and α=0.3, with no operating point in between.

E SPLIT-SEED AND COST-CONSTANT SENSITIVITY

Why this appendix exists. Downstream of the frozen response matrix the pipeline is deterministic, so the 50/50 calibration/test partition of question ids is its only source of randomness, and a single draw of it stands behind every number a reader would otherwise take as fixed. Because the whole pipeline (scores, thresholds, routing, frontier, fenced audits) re-runs on CPU in seconds, we re-derive every headline quantity under 9 split seeds: the published one and 8 consecutive offsets of it, fixed before any result was inspected so the sweep cannot select favourable splits. The sweep is a single-writer script (`split_sensitivity.py` → `results/split_sensitivity.json`) whose published-seed pass is asserted in code to reproduce the committed `frontier.json` exactly, its pool-drop pass the committed RQ4 operating points, and its baselines pass the committed `baselines.json`, so it extends the shipped run rather than re-implementing it. Table 6 reports, per audited quantity, the published-split value and the min/median/max over all 9 splits.

What this sweep does and does not vary matters for how its ranges should be read, so we state it rather than leave it to be inferred. The 9 seeds resample the calibration/test partition and nothing else: the response matrix is frozen, so item sampling and model outputs contribute exactly zero variance, and the draws are dependent by construction: any two of the 9 test folds share 47.9–52.2% of their items (measured over all seed pairs; the partition is drawn per benchmark, as in the pipeline). A min–max range here is therefore a *fold-assignment sensitivity* range conditional on this matrix, not a confidence interval, and a direction that holds on all 9 splits is not 9 independent successes; the effective sample size is well below 9, and we quote no p-value from these counts. The published split is itself one of the 9, so every range contains it. A genuine interval on a single quantity would need an item-level paired bootstrap within each split; we report the cross-seed scatter instead of one, and label it as scatter.

What survives the resampling and what does not. The paper's qualitative claims do: the certified-risk violation count is 23–25 of 30 across splits, the full probe's pooled AUROC stays in 0.554–0.572, pooled routed accuracy at α=0.10 in 68.3–69.4%, and pooled cost at the same budget in 3.75–3.98 s/item, below the CP-Router variant rerun under the identical splits (Table 7), which prices at 4.00–4.05 s/item, so CREDO undercuts it on every paired split, by 0.053 to 0.280 s/item. No cost saving against the strongest tier exists to resplit: with the probe charged, the max-over-grid matched-or-better saving is negative on every seed, between -160.0% and -157.5%, and the (benchmark, α) cell attaining it moves between seeds (the formerly published cell, HallusionBench at α=0.2, stays matched-or-better in 2 of 9 splits, passes its own certified-risk audit in 0, and swings from -214.9% to -174.8%). An audit-passing point that costs less than the strongest tier exists on 0

Table 6: Cross-split sensitivity of every audited quantity (generated from `results/split_sensitivity.json`; no hand-typed numbers). Savings are relative to the always-strongest-tier baseline; positive is cheaper. "Max. matched-or-better saving" applies the headline rule (maximum saving over the (benchmark, α) grid subject to routed accuracy $\geq$ the strongest tier's) per seed; the indented row additionally requires the point to pass its own RQ2 certified-risk audit.

Quantity	Published	Min	Median	Max
Pooled mean cost @ α=0.05 (s/item)	3.84	3.84	3.95	4.08
Pooled mean cost @ α=0.10 (s/item)	3.75	3.75	3.86	3.98
Pooled mean cost @ α=0.20 (s/item)	3.62	3.62	3.70	3.80
Pooled mean cost @ α=0.30 (s/item)	3.53	3.53	3.60	3.66
Pooled routed acc. @ α=0.10 (%)	69.0	68.3	69.0	69.4
Pooled certified risk @ α=0.10	0.242	0.206	0.225	0.247
Benchmarks meeting target @ α=0.10 (of 6)	1	0	1	1
Certified-risk violations (of 30)	23	23	24	25
Full-probe pooled AUROC	0.5627	0.5542	0.5604	0.5712
Max. matched-or-better saving (%)	-157.5	-160.0	-157.9	-157.5
restricted to audit-clean points (%)	-157.52	-160.04	-157.93	-157.52
HallusionBench@0.2 saving (%)	-176.1	-214.9	-205.4	-174.8
HallusionBench@0.2 acc. − strongest (pp)	+0.9	-3.2	-1.4	+0.9
OCRBench@0.05 saving (%)	-175.2	-176.7	-174.7	-171.2
OCRBench@0.05 acc. − strongest (pp)	+5.8	+4.8	+6.4	+8.0
OCRBench@0.10 saving (%)	-172.5	-173.4	-171.1	-166.6

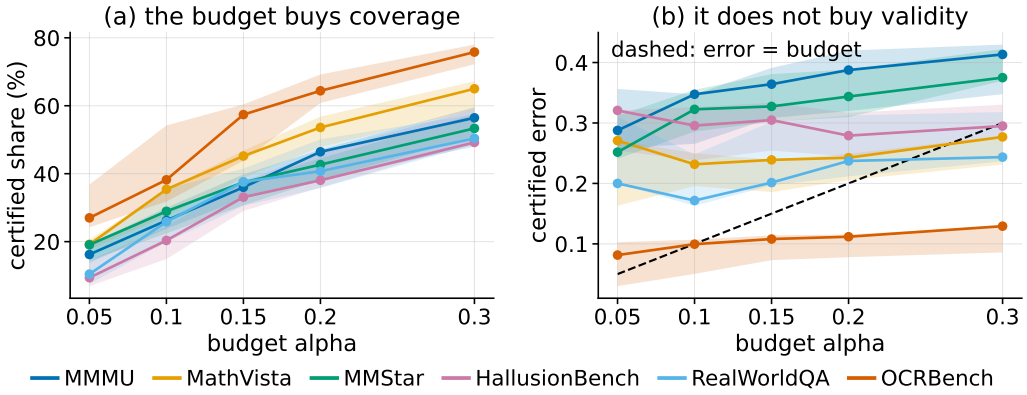

Figure 7: **What the budget buys, per benchmark and split.** Against α: (a) certified share of routed traffic and (b) conditional error of certified traffic. Line: published split; band: min–max over the 9 calibration/test splits (a fold-assignment scatter, not an interval).

of 9 splits. The one split-stable frontier gain is OCRBench at α=0.05: routed accuracy exceeds the strongest tier by +4.8 to +8.0 points on all 9 splits, at a cost between -176.7% and -171.2% relative to it; but against the best fixed tier of the same test fold, the 7B model, the gain is -1.0 to +1.2 (Appendix F), and the point clears its own certified-risk budget on only 2 of 9 splits; the audit-clean OCRBench point is α=0.1, passing on 8 of 9 splits at cost parity or worse (-173.4% to -166.6%). This is the evidence behind the main text's decision to report neither a cost nor an accuracy headline.

Baselines under the same splits. The learned baselines of Table 1 are re-run under the identical 9 splits (same seeds, same partition function, knobs re-tuned on each split's calibration fold), and Table 7 reports each family's pooled cost/accuracy spread. Families with seeded training randomness draw it from the split seed, exactly as the published run does, so their spread bundles three things and not two: the split draw, the training draw, and the in-process knob nondeterminism measured in Appendix A, whose widest family moves by 0.015 s/item on repeat draws alone. CP-Router is deterministic given the fold, so its spread is pure split spread.

The certificate axis runs the other way. The cost comparison above is one axis, and on the other axis the baseline wins on every split; reporting only the first would be selective. CP-Router's

certified-branch risk at α=0.10 spans 0.132–0.170 across splits against CREDO's 0.206–0.247 on the identical folds (Table 6), which makes CP-Router's certified branch the more valid of the two on every split. It meets its target on 1 of six benchmarks on 8 of the 9 splits and 2 on the remaining one, against CREDO's 0–1; both counts are point estimates whose borderline cells can flip by one answer (Table 1), which is most of what separates their upper ends. The two quantities are computed by the same definition (pooled conditional error of the certified subset, and the number of benchmarks whose certified risk clears the budget) on the same six benchmarks at the same α, so they are directly comparable. This is Section 4.3's base-rate gap seen from the baseline's side: CP-Router certifies a smaller, cleaner branch and charges 0.053–0.280 s/item more for it, CREDO certifies a larger one more cheaply and pays in conditional error. Neither dominates; they trade off on different axes, and the RQ2 reading of neither certificate hinges on the published split. What CREDO does get for free is quality: its routed accuracy (68.3–69.4%) exceeds CP-Router's (68.0–68.9%) on every paired split, by +0.22 to +0.92 pp: the cost gap comes without giving up routed accuracy.

Where the cheap band actually sits. Among the uncertified families the cost ordering runs against CREDO, not for it: six of the eight uncertified families in the resplit are strictly cheaper than CREDO on every split, and the RouteLLM, RouterDC and VLRouterBench cost ranges overlap CREDO's pooled α=0.10 band. That verdict is a knife-edge, not a margin: CREDO's own cross-split standard deviation is 0.077, and AutoMix, at 4.13 s/item on the published split, sits entirely above the band, its cross-split maximum -0.375 s/item short of CREDO's minimum. A min–max overlap test on 9 dependent draws is not a finding to lean on, and we do not: the paper's cost claim against the uncertified band is the one in Section 4.2, that they are cheaper and carry no certificate.

Table 7: Learned baselines under the same 9 calibration/test splits (generated from the per-seed baselines section of `results/split_sensitivity.json`; no hand-typed numbers). Pooled test-fold mean cost and routed accuracy; the published-split column matches Table 1 and is one of the 9 splits the min/median/max are taken over, and the published-split pass is asserted in code to reproduce the committed `baselines.json`. †/‡ as in Table 1; § marks the RouteLLM scorer variant Table 1 does not carry (Appendix C), included here so the cheap-band statement above can be audited over the same eight uncertified families it is scored on.

	Cost (s/item)				Routed acc. (%)			
Baseline	Pub.	Min	Med.	Max	Pub.	Min	Med.	Max
CP-Router (pairwise)†	4.02	4.00	4.02	4.05	68.8	68.0	68.2	68.9
kNN router	3.63	3.52	3.63	3.67	67.2	66.0	67.5	69.4
AutoMix†	4.13	4.04	4.10	4.13	69.5	68.2	69.1	69.5
RouteLLM transplant (logistic)	3.79	3.72	3.83	3.92	68.0	67.4	68.0	69.1
RouteLLM transplant (grad.-boosted)§	3.20	3.20	3.21	3.23	63.7	62.0	63.5	64.1
ECVL-Router (variant)‡	3.48	3.44	3.50	3.56	66.6	66.2	67.0	68.4
LatentRouter (variant)‡	3.19	3.16	3.19	3.20	63.7	61.7	63.7	65.1
VL-RouterBench ref. (variant)‡	3.64	3.53	3.61	3.76	67.5	67.0	67.9	69.2
RouterDC (variant)‡	3.67	3.66	3.71	3.77	67.5	67.5	68.4	69.1

The tier-1 cost constant. Every cost number charges each query the smoke-measured sec/item of the tier that answers it. The tier-1 constant 3.07 s/item is conservative: its measurement window bundles two OOM-aborted window attempts and four engine starts, while the clean final smoke cell ran 2.04 s/item (the ledger's final smoke entry cross-checks the clean figure to its rounding, but is not independently re-derivable from the committed ledger alone, which is why 3.07 remains the shipped constant and 2.04 is reported only as a sensitivity). Tier costs never enter a routing decision (the cascade routes on certificates and the mini-max rule on risk intervals), so the identical decisions reprice exactly under either constant. Repricing shifts magnitudes and no directions: the published-split HallusionBench α=0.2 saving moves from -174.8% to -94.3%, and the RQ4 full-pool operating point from 3.75 to 2.72 s/item with per-drop shifts spanning -9.9% to +6.2% (shipped: -7.2% to +4.5%). Tier 1 remains the pool's most expensive tier under the clean constant as well, so the cost-order inversion, the fact the design-time cascade assumption fails on, is not an artifact of the contaminated window. Re-running the tier-1 smoke three times on a second machine (RTX 4090D, batch 1) reproduces the frozen prediction string on 170 of 180 rows (94.4%, the same ten rows differing each time) at 1.32–1.36 s/item; the constant is machine-specific and the footnote stands.

Rebuilding the response matrix. The `matrix_path` field in `frontier.json`, `ablation.json`, `baselines.json` and `split_sensitivity.json` records only the *file name* of the matrix each artifact was computed from, with the directory written as `<scratch>/`: the producing session's directory was neither portable nor anonymous, so the writers record a repo-relative path when the input lives in the repository and the bare file name otherwise. Nothing is lost, because the matrix is not distributed as that single file. It is released per cell under `results/raw/<model>/<benchmark>/responses.jsonl` and is rebuilt by concatenation: `cat results/raw/*/*/responses.jsonl`. `split_sensitivity.json` carries that rule in its `matrix_rebuild_rule` field, and its published-seed no-drift assertion against the committed frontier and ablation is the external check that a matrix rebuilt this way is the matrix the paper's numbers came from.

F BEST FIXED TIER, CALIBRATION-FOLD PREDICTION, AND THE ABLATION TABLE

Everything in this appendix is generated by `experiments/credo/deep_analysis.py` from the released matrix and artifacts; its published-seed pass is asserted to reproduce `frontier.json`, and every per-split routed-minus-strongest delta to reproduce `split_sensitivity.json`, before it writes anything.

Best fixed tier per split. The strongest tier by design is Qwen2.5-VL-32B-AWQ, but the best *fixed* choice on a test fold is whichever single tier scores highest there, chosen with test labels, an optimistic baseline, which is intended. Table 8 gives, per benchmark and budget, routed accuracy minus that tier's accuracy on the published split and its min/max over the 9 splits. The best tier is not the 32B model on two of the six benchmarks of the published split and on 20 of the 54 (split, benchmark) folds. Over all 270 combinations the largest gain is +1.6 points (OCRBench at α=0.2) and 0 reach 3 points; routed accuracy is at least the best fixed tier's at 43 of them (5 of 30 on the published split).

Table 8: Routed accuracy minus the best fixed tier's test-fold accuracy (pp), per benchmark and budget: published split (pub) and min/max over the 9 splits (generated). "Best tier" is the published split's.

Benchmark	Best tier	α=0.05			α=0.1			α=0.15			α=0.2			α=0.3		
		pub	min	max	pub	min	max	pub	min	max	pub	min	max	pub	min	max
MMMU	32B	-0.9	-2.5	-0.2	-0.4	-2.5	+0.3	-1.1	-2.9	+0.7	-1.6	-4.9	+0.9	-2.7	-6.0	-0.6
MathVista	32B	-3.0	-4.8	-0.2	-3.2	-5.2	-1.4	-4.0	-6.4	-0.6	-3.6	-5.2	-0.8	-4.4	-5.8	-1.2
MMStar	32B	-0.7	-1.1	+0.3	-1.9	-2.0	-0.5	-2.4	-3.2	-0.8	-3.5	-4.6	-1.6	-4.7	-5.4	-3.4
HallusionBench	7B	-1.1	-3.8	-0.3	-1.1	-3.9	+0.0	-1.1	-4.8	-0.3	+0.2	-4.8	+0.2	-1.2	-6.2	-1.2
RealWorldQA	32B	-0.8	-4.0	+0.3	-0.3	-3.2	+1.4	-1.0	-4.7	+0.8	-2.3	-5.8	+0.3	-3.1	-6.8	-2.6
OCRBench	7B	+0.4	-1.0	+1.2	+0.4	-1.4	+0.8	+0.2	-1.6	+1.0	+0.2	-2.2	+1.6	-0.2	-3.0	+1.4

Calibration-fold prediction of the certified error. Table 9 lists, per (benchmark, α) on the published split, the certified-traffic error of the cascade replayed on the calibration fold (the prediction), the test-fold value (the measurement), the calibration fold's chance limit $\bar{\pi}$ (the certified-traffic-weighted tier wrong rate), and the two violation verdicts. Certified-subset sizes are in parentheses.

Ablation table. Table 10 is the table form of Figure 5. Run alone, the four channels pool to logprob/entropy 0.548, vote entropy 0.538, question features 0.519, and CLIP alignment 0.502 (a channel absent from a cell contributes the capacity-prior constant there).

G BUDGET ARITHMETIC

The response matrix is 6,294 questions $\times$ 4 tiers = 25,176 generations. Pre-smoke estimates total $\sim$15 GPU-hours ($\sim$33 with a 2$\times$ contingency and the bf16 fallback delta), against the 64 GPU-hour cap; the remainder was reserved for the in-pool baseline self-runs, which in the end cost zero GPU-hours, since every reproducible baseline is a CPU rescan of the same matrix (Appendix C). All per-tier sec/item estimates are marked `smoke_pending` and are overwritten by the measured

Table 9: Calibration-fold prediction versus test-fold measurement of the certified-traffic error, published split (generated).

Benchmark	α	Predicted (cal. fold)	Measured (test fold)	$\bar{\pi}$ (cal. fold)	Verdict pred. / meas.
MMMU	0.05	0.270 (74)	0.288 (73)	0.516	viol. / viol.
	0.1	0.339 (109)	0.347 (118)	0.517	viol. / viol.
	0.15	0.364 (151)	0.364 (162)	0.519	viol. / viol.
	0.2	0.347 (193)	0.388 (209)	0.521	viol. / viol.
	0.3	0.404 (240)	0.413 (254)	0.524	viol. / viol.
MathVista	0.05	0.167 (102)	0.271 (96)	0.388	viol. / viol.
	0.1	0.190 (179)	0.232 (177)	0.385	viol. / viol.
	0.15	0.211 (218)	0.239 (226)	0.388	viol. / viol.
	0.2	0.218 (261)	0.243 (268)	0.391	viol. / viol.
	0.3	0.247 (320)	0.277 (325)	0.393	ok / ok
MMStar	0.05	0.324 (111)	0.252 (143)	0.448	viol. / viol.
	0.1	0.321 (190)	0.323 (217)	0.450	viol. / viol.
	0.15	0.363 (256)	0.327 (281)	0.449	viol. / viol.
	0.2	0.383 (303)	0.344 (320)	0.450	viol. / viol.
	0.3	0.420 (386)	0.375 (400)	0.451	viol. / viol.
HallusionBench	0.05	0.268 (56)	0.321 (53)	0.345	viol. / viol.
	0.1	0.292 (113)	0.296 (115)	0.345	viol. / viol.
	0.15	0.267 (187)	0.305 (187)	0.330	viol. / viol.
	0.2	0.289 (232)	0.279 (215)	0.324	viol. / viol.
	0.3	0.316 (297)	0.295 (278)	0.325	viol. / ok
RealWorldQA	0.05	0.294 (34)	0.200 (40)	0.387	viol. / viol.
	0.1	0.262 (84)	0.172 (99)	0.384	viol. / viol.
	0.15	0.234 (128)	0.201 (144)	0.381	viol. / viol.
	0.2	0.280 (150)	0.237 (156)	0.381	viol. / viol.
	0.3	0.308 (185)	0.244 (193)	0.385	viol. / ok
OCRBench	0.05	0.027 (146)	0.081 (135)	0.162	ok / viol.
	0.1	0.052 (191)	0.099 (191)	0.158	ok / ok
	0.15	0.058 (294)	0.108 (287)	0.162	ok / ok
	0.2	0.073 (329)	0.112 (322)	0.164	ok / ok
	0.3	0.087 (381)	0.129 (379)	0.161	ok / ok

30-item smoke run before any full launch; the in-code launch guard re-checks the ledger against whatever estimates are committed at that moment.

H PROBE TIERS BEYOND 3B (PRE-REGISTERED EXTENSIONS B2, B3)

Registered in `PREREG_PROBE_TIERS.md` before any row. The probe channels were regenerated with the 7B tier (k=5 self-consistency and top-20 logprobs) and the 32B-AWQ tier (greedy and logprobs) on the frozen matrix's questions, batch 1. Each run's greedy prediction agreed with the frozen row on 87.8% (7B, 5527/6294; MathVista 49.0%) and 96.6% (32B) of rows, below the registered 0.97 bar, so channel scores are computed on the agreeing rows. Pooled over the ablation's cells, the logprob/vote channels reach 0.551 AUROC at 7B (0.549 without the rule) against 0.560 (0.550) at 3B; single channels 0.547/0.532 against 0.548/0.538. The registered reading (within $+0.05$: probe capacity does not set the AUROC) holds. Each tier's confidence in its own answers (logprob self-AUROC on the test fold) is 0.696 at 3B, 0.702 at 7B and 0.689 at 32B-AWQ (single-logp pooled 0.545 against 0.548); the strongest tier's self-confidence is the lowest of the three, between the two registered branches, and is reported as measured.

I A SEVENTH, DESIGN-EXTERNAL BENCHMARK (PRE-REGISTERED EXTENSION B5)

Registered in `PREREG_B5_SEVENTH_BENCHMARK.md` before any row; ChartQA was excluded before generation because its official relaxed-accuracy metric is not implemented in the repository's

Table 10: **RQ3: probe-signal rescans** (leave-one-out; single-channel values in the text). AUROC is test-fold correctness discrimination per (benchmark, tier) cell, pooled by test-fold size; cost is the certified frontier's pooled mean at α=0.10. All rows are full-pipeline rescans of the cached matrix on identical folds; leave-one-out deltas are within split-resampling scatter.

Probe variant	AUROC	Mean cost @ α=0.10 (s/item)
Full probe (all channels)	0.563	3.75
− logprob / entropy	0.559	3.74
− vote entropy	0.554	3.78
− CLIP alignment	0.560	3.76
− question features	0.549	3.72

scorers, and AI2D (official metric: option-letter exact match) was used: 3088 test questions, the four reported tiers, greedy, batch 1 (tier accuracies 0.776, 0.818, 0.806, 0.831). Two evaluations (Table 11) over the same nine question-disjoint splits: *transfer*, the thresholds calibrated on the six design benchmarks applied unchanged; *own-split*, the shipped calibration on the seventh benchmark's own half. Transferred thresholds hold the risk at 4 of 5 α points (the exception α=0.10: 0.119), which under the registered rule reads as the certificate transferring; own-split calibration holds at 4 of 5. Own-split coverage at α=0.10 is 47.3% against the seed-mean over the six design benchmarks, 28.7% (Table 1 pools the published split instead), +18.7 pp outside the registered $\pm$15 pp band; the benchmark is easier than the design pool.

Table 11: Seventh benchmark (AI2D): seed-mean certified risk and coverage per α under transferred and own-split thresholds, with the six-benchmark mean coverage.

α	transfer risk	transfer cov.	own-split risk	own-split cov.	six-benchmark cov.
0.05	0.017	0.080	0.063	0.302	0.176
0.1	0.119	0.457	0.076	0.473	0.287
0.15	0.123	0.634	0.093	0.602	0.395
0.2	0.160	0.764	0.110	0.681	0.470
0.3	0.174	0.888	0.130	0.768	0.587

J PER-BENCHMARK FRONTIER TABLES

Table 12 breaks Table 1 down per benchmark; it is generated directly from `results/frontier.json` by the macro bridge (no hand-typed numbers). "Abstain (policy)" is the abstain-policy counterpart computed on the same folds; the default strongest-fallback policy answers every query.

K A FIFTH TIER FROM A THIRD MODEL FAMILY (PRE-REGISTERED EXTENSION B1)

Registration and candidate order. The main pool is three Qwen2.5-VL sizes plus LLaVA-OneVision-7B, so every language backbone in it traces to Qwen2. Whether the validity finding of §4.3 is an artifact of that make-up was registered as item B1 of `PREREG_TIER5_PROBE.md` (committed 2026-09-10, before any fifth-tier row existed; seven same-evening amendments, all pre-run; the file ships in the supplementary archive). The candidate order was fixed in advance: Idefics3-8B-Llama3, Phi-3.5-vision-instruct, pixtral-12b; the first candidate to pass Step 0 is the tier. Step 0 requires the architecture in the vLLM 0.7.3 registry, loading under the pinned environment (transformers 4.49.0) with no change to what the harness computes, a 30-item greedy smoke per benchmark parsing at least 0.9 under each official scorer, and, strengthened after candidate 1's failure, clean generation of every multi-image MMMU item as a second stratum (43 items). Table 13 records what each candidate did; the machine-readable record is `results/step0_tier5.json` and the run scripts, written for a lab workstation, are under `runbook/tier5/` in the supplementary archive. The chosen tier parses at 0.9666 or better on every benchmark.

Table 12: Per-benchmark certified operating points (generated). **Bold** certified risk marks a violation (risk $> \alpha$).

Benchmark	α	Cost (s/item)	Routed acc. (%)	Cert. share	Cert. risk (n)	Abstain (policy)
MMMU	0.05	4.15	56.2	0.16	**0.288** (73)	0.84
	0.1	4.05	56.7	0.26	**0.347** (118)	0.74
	0.15	3.93	56.0	0.36	**0.364** (162)	0.64
	0.2	3.81	55.6	0.46	**0.388** (209)	0.54
	0.3	3.70	54.4	0.56	**0.413** (254)	0.44
MathVista	0.05	3.52	66.8	0.19	**0.271** (96)	0.81
	0.1	3.47	66.6	0.35	**0.232** (177)	0.65
	0.15	3.45	65.8	0.45	**0.239** (226)	0.55
	0.2	3.40	66.2	0.54	**0.243** (268)	0.46
	0.3	3.33	65.4	0.65	0.277 (325)	0.35
MMStar	0.05	4.12	66.7	0.19	**0.252** (143)	0.81
	0.1	4.01	65.5	0.29	**0.323** (217)	0.71
	0.15	3.91	64.9	0.37	**0.327** (281)	0.63
	0.2	3.84	63.9	0.43	**0.344** (320)	0.57
	0.3	3.72	62.7	0.53	**0.375** (400)	0.47
HallusionBench	0.05	3.57	70.3	0.09	**0.321** (53)	0.91
	0.1	3.52	70.3	0.20	**0.296** (115)	0.80
	0.15	3.51	70.3	0.33	**0.305** (187)	0.67
	0.2	3.48	71.5	0.38	**0.279** (215)	0.62
	0.3	3.43	70.1	0.49	0.295 (278)	0.51
RealWorldQA	0.05	4.22	70.8	0.10	**0.200** (40)	0.90
	0.1	4.04	71.3	0.26	**0.172** (99)	0.74
	0.15	3.91	70.5	0.38	**0.201** (144)	0.62
	0.2	3.87	69.2	0.41	**0.237** (156)	0.59
	0.3	3.75	68.4	0.50	0.244 (193)	0.50
OCRBench	0.05	3.47	84.8	0.27	**0.081** (135)	0.73
	0.1	3.43	84.8	0.38	0.099 (191)	0.62
	0.15	3.32	84.6	0.57	0.108 (287)	0.43
	0.2	3.29	84.6	0.64	0.112 (322)	0.36
	0.3	3.24	84.2	0.76	0.129 (379)	0.24

How the fifth tier was served. The four published tiers were generated at batch 16; the fifth at batch 1. Batch size is serving configuration (prompts, decoding and scorer are untouched), and this stack does not give string equality across batch sizes: on candidate 1's 30-item smoke, 5 of 180 predictions differ between the two batch sizes while a repeat run at a fixed batch size reproduces every one, and the official scorers move 80 to 79 correct of 150 scorable items (3 label flips; `results/batch_invariance_tier5.json`, `results/batch_scoring_tier5.json`). What the matrix needs is that the tier's generation is one reproducible computation, so the registered gate is self-consistency at the batch size actually used: the batch-1 smoke reproduces the full run's first 30 rows per benchmark string-for-string on all six benchmarks (`results/self_consistency_tier5.json`). The fifth tier's cells live in `results/raw_tier5/`, outside the four-tier rebuild rule; the five-tier matrix assembly is recorded in `results/matrix5_tier5.manifest.json`.

T1: pool dependence of the validity finding. The registered decision rule: if the five-tier count differs from the four-tier count by more than two operating points, the finding is pool-dependent and the abstract's sentence is scoped to the four-tier pool. The five-tier count is 23 of 30 on the published split against a recomputed four-tier baseline of 23 (+0), and runs 23–25 across the 9 split seeds (Table 14): within the band, so adding a tier from a third family leaves the count within two operating points and the abstract stands unscoped.

T2: the per-tier bound on the new tier. The wrong-subset construction bounds P(accept | wrong) by α per tier; Table 15 measures that rate for the new tier on the test fold. Per cell the new tier straddles α as the published four do: 14 of 30 (benchmark, α) cells above at the published split,

Table 13: B1 Step 0: the registered candidate order and outcomes (generated from `results/tier5_readout.json`). "Registry" is membership in the vLLM 0.7.3 model registry; "loads" is the pinned stack serving the checkpoint without a harness change; "worst parse" is the lowest official-scorer parse rate over the six 30-item smokes. Candidate 3's checkpoint declares `LlavaForConditionalGeneration` and is served through vLLM's LLaVA class with a Pixtral vision tower.

#	candidate	registry	loads	worst parse	outcome
1	`idefics3-8b-llama3`	yes	yes	0.9333 (MathVista)	failed at full-run time: vLLM 0.7.3 Idefics3 cannot batch differing image counts (MMMU row 49) nor one request with two differently shaped images at batch 1 (row 84); Amendment 3 records it as failing Step 0's runs-under-pinned-stack condition; no fifth-tier row exists
2	`phi-3.5-vision`	yes	no	not reached	failed Step 0: the pinned transformers 4.49.0 Phi3VProcessor has no chat_template attribute, so the harness list-of-parts messages do not render without a code change (Amendment 3's named failure class)
3	`pixtral-12b`	yes	yes	0.9667 (MathVista)	chosen

Table 14: T1: certified-subset validity violations on the five-tier pool per split seed, with the recomputed four-tier baseline (generated). The published seed is one of the nine.

split seed	five-tier violations (of 30)
four-tier baseline (recomputed)	23
20260718 (published)	23
20260719	24
20260720	24
20260721	23
20260722	24
20260723	25
20260724	25
20260725	23
20260726	24

6–20 across the 9 seeds. Marginally (the quantity the bound states), the pooled rate over its 1345 wrong test answers is 0.048, 0.119, 0.158, 0.207 and 0.297 at α=0.05, …, 0.3. Under the rule Section 4.3 applies to the four published tiers (above budget when a one-sided exact binomial test of rate $\leq \alpha$ rejects at level 0.05), it is above budget at 1 of the five levels (α=0.10, p=0.013) and at budget at the rest: a violation of the per-tier bound on the new tier at that level, reported here per the registration, at the prominence of the per-tier claim.

Table 15: T2: measured P(accept | wrong) on the fifth tier, per (benchmark, α) cell, published split (generated). **Bold** exceeds α. The pooled row is the quantity the marginal bound states.

benchmark	wrong test	α=0.05	α=0.10	α=0.15	α=0.20	α=0.30
HallusionBench	196	0.046	0.066	0.092	0.138	0.240
MathVista	217	0.046	**0.143**	**0.198**	**0.253**	**0.359**
MMMU	227	0.048	**0.176**	**0.225**	**0.300**	**0.379**
MMStar	384	0.049	**0.117**	0.135	0.167	0.242
OCRBench	183	**0.060**	**0.104**	**0.175**	**0.224**	**0.311**
RealWorldQA	138	0.029	0.087	0.123	0.174	0.275
pooled	1345	0.048	**0.119**	**0.158**	**0.207**	0.297

T3: a different tier, or a re-measured one? Table 16 pools the Pearson correlation of per-question correctness over all questions, for every pool pair; Table 17 gives the test-fold accuracies those correlations are taken over. The new tier tracks each published tier at r=0.324–0.387, the bottom of the range the published pairs span (0.357–0.515), and below every Qwen-internal pair. No threshold was registered for T3; descriptively, the pool gained a family rather than a duplicate.

Table 16: T3: Pearson correlation of per-question correctness, pooled over all six benchmarks (generated). Tier short names: 3B/7B/32B = Qwen2.5-VL, OV-7B = LLaVA-OneVision-7B, PX-12B = pixtral-12b.

tier pair	pooled r
3B vs 7B	0.515
32B vs 7B	0.488
32B vs 3B	0.431
OV-7B vs 32B	0.387
OV-7B vs PX-12B	0.386
PX-12B vs 32B	0.386
OV-7B vs 7B	0.372
PX-12B vs 7B	0.362
OV-7B vs 3B	0.357
PX-12B vs 3B	0.324

Table 17: Test-fold accuracy of each pool member per benchmark on the published split, the pool Table 16 is taken over (generated).

tier	HallusionBench	MathVista	MMMU	MMStar	OCRBench	RealWorldQA
3B	0.665	0.586	0.478	0.565	0.808	0.611
7B	0.713	0.682	0.467	0.639	0.844	0.687
OV-7B	0.609	0.618	0.467	0.639	0.612	0.674
32B	0.706	0.698	0.571	0.673	0.790	0.715
PX-12B	0.653	0.566	0.496	0.488	0.634	0.640

L A HIDDEN-STATE DIFFICULTY PROBE ON THE 3B TIER (PRE-REGISTERED EXTENSION B4)

Registration: `PREREG_TIER5_PROBE.md`, item B4; read-out `experiments/credo/results/READOUT_probe.md`. One prefill of the pool's tier 1 (`Qwen2.5-VL-3B-Instruct`, pinned revision) on the prompt the matrix used, no generation, last-layer hidden states pooled two ways (mean over prompt tokens; final prompt token), 0.56 GPU-h for 6,294 questions. Per tier a logistic-regression probe from the pooled vector to the official scorer's correctness is trained on one half of the calibration fold and scores the other half and the test split; nothing else in the certificate changes. Features come from a transformers-side prefill, which reproduces the serving engine's first token on 282 of 300 verification items (gate 0.9); the top-k logprob gap is reported, not gated (median 0.1422, max 1.1213). Fed the shipped rule probe, the same runner reproduces the committed frontier at every operating point (23 of 30 over budget); halving the calibration fold alone gives 24. The registered readings (Table 18): H1 holds for a pooling if at least 4 of six benchmarks reach AUROC 0.70 on the 7B tier; H2 is supported if the count falls to at most 15, and reads "over-admission is not probe-limited on this pool" at 20 or more.

Table 18: B4 read-out, published split; the seed range spans the nine split seeds. The certified share is the min–max over the 30 operating points.

score	pooled AUROC	benchmarks at bar (7B)	over budget (seed range)	certified share (%)
rule probe, full calibration fold	0.563	—	23	9.3–75.8
rule probe, halved calibration fold	—	—	24	—
hidden state, mean pooling	0.629	1	26 (23–27)	13.5–69.1
hidden state, last pooling	0.718	4	20 (20–25)	19.8–77.0

M A LEARNED FUSION OF THE PROBE'S INPUTS (EXTENSION A10)

Read-out `experiments/credo/results/READOUT_learned_probe.md`. The rule probe pools its 16 per-question inputs (12 question-surface features, the CLIP score, and the 3B tier's logprob, entropy and vote entropy) with fixed weights. This extension asks how much of the RQ3 null is the pooling: the same inputs, with a benchmark one-hot for the pooled arms, are fed to a logistic or gradient-boosted model fit on one half of the calibration fold and scored on the published test fold,

over the 9 split seeds of Appendix E; the conformal quantile is then taken on the held-out calibration half and the 30 operating points are re-audited. Table 19 gives the pooled and per-benchmark AUROC (mean and seed range) for the four learned arms against the shipped rule, then the attribution arms: the 3B channels alone, with CLIP, question-surface features alone, and a permuted-label control. The per-benchmark logistic arm is the best at 0.679 pooled; the 3B channels alone give 0.661, the surface features alone 0.585, and the permuted control 0.494. The certificate on the best arm is over budget at 21–23 of 30 points across seeds, against 24 for the rule probe on the same halved fold. The signal the rule leaves on the table is real; the over-admission is not.

Table 19: A10 read-out: pooled and per-benchmark test AUROC, mean over the 9 split seeds with the seed range beneath. Top: the learned arms against the shipped rule. Bottom: attribution arms of the per-benchmark logistic model and the permuted-label control.

BENCHMARK	RULE SHIPPED	LR POOLED	LR PER-BENCH.	GBM POOLED	GBM PER-BENCH.
Pooled	0.561 [0.554, 0.572]	0.662 [0.654, 0.674]	0.679 [0.671, 0.689]	0.660 [0.650, 0.669]	0.648 [0.638, 0.663]
HallusionBench	0.551 [0.531, 0.573]	0.596 [0.573, 0.626]	0.621 [0.593, 0.639]	0.609 [0.597, 0.622]	0.586 [0.563, 0.602]
MathVista	0.628 [0.617, 0.642]	0.691 [0.683, 0.707]	0.704 [0.684, 0.716]	0.696 [0.668, 0.719]	0.702 [0.684, 0.727]
MMMU	0.544 [0.514, 0.567]	0.692 [0.664, 0.707]	0.695 [0.667, 0.718]	0.655 [0.629, 0.667]	0.632 [0.592, 0.654]
MMStar	0.511 [0.496, 0.537]	0.661 [0.652, 0.674]	0.668 [0.657, 0.679]	0.657 [0.643, 0.669]	0.647 [0.634, 0.656]
OCRBench	0.600 [0.580, 0.612]	0.683 [0.665, 0.715]	0.736 [0.715, 0.761]	0.719 [0.667, 0.747]	0.710 [0.688, 0.747]
RealWorldQA	0.557 [0.529, 0.583]	0.661 [0.640, 0.677]	0.660 [0.625, 0.688]	0.628 [0.597, 0.667]	0.608 [0.575, 0.636]

BENCHMARK	RULE 3B, ALL TIERS	LR 3B ONLY	LR 3B + CLIP	LR SURFACE ONLY	LR PERMUTED
Pooled	0.635 [0.628, 0.646]	0.661 [0.657, 0.672]	0.662 [0.652, 0.674]	0.585 [0.574, 0.594]	0.494 [0.473, 0.527]
HallusionBench	0.561 [0.540, 0.594]	0.625 [0.597, 0.644]	0.628 [0.594, 0.659]	0.573 [0.563, 0.587]	0.495 [0.466, 0.543]
MathVista	0.670 [0.655, 0.693]	0.663 [0.647, 0.686]	0.666 [0.656, 0.683]	0.630 [0.599, 0.652]	0.493 [0.414, 0.594]
MMMU	0.674 [0.649, 0.692]	0.706 [0.689, 0.736]	0.703 [0.679, 0.737]	0.558 [0.525, 0.575]	0.459 [0.357, 0.522]
MMStar	0.612 [0.591, 0.633]	0.664 [0.651, 0.672]	0.661 [0.650, 0.672]	0.558 [0.542, 0.575]	0.497 [0.433, 0.536]
OCRBench	0.686 [0.675, 0.714]	0.655 [0.638, 0.670]	0.661 [0.650, 0.681]	0.633 [0.592, 0.661]	0.520 [0.463, 0.584]
RealWorldQA	0.629 [0.607, 0.657]	0.666 [0.645, 0.679]	0.664 [0.641, 0.679]	0.563 [0.535, 0.601]	0.491 [0.442, 0.553]