TWENTY LOGPROBS ARE NOT ENOUGH: THE FEASIBILITY FRONTIER FOR CONFORMAL PREDICTION UNDER BLACK-BOX OUTPUT TRUNCATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction assumes the analyst can score every candidate label. Production APIs violate it: the interfaces we surveyed return at most 5–50 log-probabilities per position, some none. We characterize what distribution-free coverage survives this truncation. Coverage is capped at top-k accuracy plus a *tail-concentration* term we define; for rules that neither single out hidden labels nor reach past the returned window the bound reads $\alpha \geq e_k$: the interface, not the score, limits confidence. A matching truncated procedure attains the bound marginally (not given a usable set), and a binomial criterion with a phase transition at $e_k = \alpha$ governs collapse: the population budget $k_{\min}(\alpha)$ is necessary but 1.23–1.36$\times$ too small at practical n. We verify the frontier over 670,000 scored predictions and a 49,696-call audit of seven production endpoints; two deliver less than the access level their schema advertises (one at the temperature a classifier would use, one in default reasoning mode), and returned probabilities imply censoring rates up to 7,856$\times$ below the realized ones. Where the tail has an atom the term is an escape route: one calibration-identified appended label recovers 53% offline and 74–100% on six of seven live endpoints. Feasibility must be certified from labels, not from the API's numbers. Code and records are in the supplement and at `https://github.com/xxx/xxx` (public upon acceptance).

1 INTRODUCTION

Conformal prediction wraps any classifier into a set-valued predictor with a distribution-free guarantee: for exchangeable data, $C(X_{n+1})$ contains the true label with probability at least $1 - \alpha$ (Vovk et al., 2005; Papadopoulos et al., 2002; Lei et al., 2018). Scoring every label against a calibration quantile, it is the default uncertainty recipe for deployed classifiers (Angelopoulos & Bates, 2023).

The recipe rests on an assumption rarely made explicit: *the score must be computable at every label*. For a classifier served over a network interface, it is not. Of 28 production endpoints we probed (2026-08), 17 answered and 11 of those serve usable log-probabilities at some depth: OpenAI and DeepSeek cap at 20 per token position, Alibaba's Qwen at 5, Zhipu's GLM at 50; the other six refuse informatively (Appendix B). Cloud vision services return top-5 tags, speech systems N-best lists. The probability vector every nonconformity score consumes (LAC, Sadinle et al., 2019; APS, Romano et al., 2020; RAPS, Angelopoulos et al., 2021; SAPS, Huang et al., 2024) is observable only through a keyhole; one recent method needs "the full logit outputs prior to any top-k ... truncation" (Kotla & Kotla, 2025), which no commercial API supplies. The closest prior work, Su et al. (2024), builds a conformal procedure for API-only LLMs from *sampled* responses: level L3 of §3, a method rather than a limit. Missing is the frontier: given k scores and not K, what is achievable.

This paper characterizes conformal prediction under *output truncation*: the analyst sees only the k most probable labels and their probabilities. The formal results are deliberately elementary (an exchange argument, an order statistic, a binomial criterion and a one-line Lagrangian), and the contribution is the frontier they jointly pin down and the production audit built on it.

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.

The ceiling (§4). Theorem 1 bounds the coverage of *any* rule built on the returned window, randomized and free to use the calibration labels, by $\mathrm{Acc}_k + \tau_k(b)$, where b bounds the labels predicted from outside the visible top-k and τ_k, the *tail-concentration profile*, is the most coverage appending b hidden labels could buy. For *tail-blind* rules the hidden term is at most $b\, e_k/(K-k)$, negligible when $b \ll K-k$, and the bound collapses to a clean impossibility: **certifiable coverage cannot exceed top-k accuracy**, so certifying $1-\alpha$ requires $k_{\min}(\alpha) = \min\{k : e_k \le \alpha\}$ log-probabilities. That quantity is prior art as a *construction* under full access (Shi et al., 2024, Eq. 7); the new half is necessity. The qualifier earns its place: one calibration-mined hidden label recovers 53% of the removed coverage on CLINC150 (§4), a held-out lower bound on τ_k.

The floor, and when it collapses (§5). Replacing the unobservable score by $+\infty$ is valid at every finite sample size (Proposition 1) and non-degenerate with probability $\to 1$ when $e_k < \alpha$ (Theorem 2), so the frontier $\alpha \ge e_k$ is tight. With too many censored calibration points the quantile is $+\infty$ and the procedure returns everything; Theorem 2 gives the exact law, a binomial crossing $\lfloor (n+1)\alpha \rfloor$, with a phase transition at $e_k = \alpha$ and a calibration-size requirement the population statement hides: inside the feasible region ($e_5 = 0.0712$ vs. $\alpha = 0.1$) it still degenerates with probability 0.283 at $n = 50$. The censored slice has conditional coverage identically zero (Proposition 2).

The hierarchy (§6–§7). Interfaces differ in kind: L0 (full vector), L1 (top-k probabilities), L2 (ranks only), L3 (samples only). Under L2, identity-blind rules that see only the ranked list, on a rank-calibrated model, cannot adapt: every admissible rule is dominated by a top-r set with r independent of the input, and the law of the set size cannot depend on the input (Theorem 3). Within the 20-logprob cap the rank envelope is a median $1.41\times$ the size of truncated LAC at matched coverage, never smaller: a comparison of two procedures, not a bound. When depth can be bought (tier upgrades, logit-bias peeling (Carlini et al., 2024; Finlayson et al., 2024; Morris et al., 2024), cascades), the cost-optimal per-input policy is an *absolute probability threshold*, i.e. ϵ-sampling (Hewitt et al., 2022) and not nucleus and not min-p's rescaled floor (Theorem 4). A top-k user can still audit the damage: Proposition 3 brackets the full-access threshold from top-k data.

The measurement (§9). Eight open-weight LLMs (0.5B–7.6B) at *full vocabulary* over 400,000 token positions, nine ImageNet checkpoints over 270,000 images with synthesized shift, and a 49,696-call audit of seven endpoints over five tasks. Scale helps: within one tokenizer family $k_{\min}(0.1)$ falls $3.4\times$ from 0.5B to 7.6B, and every model needs at least $50\times$ the 20-logprob cap at α=0.01. Two of seven endpoints deliver less than their schema advertises (one returns one-hot probabilities at the deterministic setting a classifier would use, so is L2 in fact; the other, in its default reasoning mode, burns 76% of its 50-logprob window on preamble), and the censoring rates the returned probabilities imply understate the realized ones by up to $7{,}856\times$. On the L2-in-fact one, split conformal at α=0.05 keeps 49.4 labels with the advertised probabilities and 7.9 with the rank score, which needs less access. In practice, *returned probabilities cannot certify a feasibility budget, but labeled calls can*, a few hundred away from the frontier and Corollary 4's count near it.

2 Related work

Conformal classification under full access. Split conformal prediction (Papadopoulos et al., 2002; Vovk et al., 2005; Lei et al., 2018) with LAC/THR (Sadinle et al., 2019), APS (Romano et al., 2020), RAPS (Angelopoulos et al., 2021) and SAPS (Huang et al., 2024) is our baseline object; all four are rank-monotone (Lemma 1), the paper's structural hinge (the rank itself is a score in Luo & Zhou, 2024), and class-conditional variants (Ding et al., 2023) sharpen the guarantee truncation destroys first. Set-valued classification studies the (coverage, expected-size) frontier and size-minimal thresholding directly (Denis & Hebiri, 2017; Chzhen et al., 2021); the achievability halves of Theorems 3 and 5 are its rank-only and truncated instances — new only in the access restriction.

Rank truncation as a design choice. RC3P (Shi et al., 2024) defines $\min\{k : e_y^k < \alpha\}$ class-conditionally and budgets top-k error out of α, trading it against the cost of scanning all labels; SAPS (Huang et al., 2024) discards every probability but the top one in favour of rank terms. Both *choose* truncation under full access. We prove the converse: under top-k *access* the same quantity is a ceiling no tail-blind procedure can beat, with an exact law for when the procedure collapses.

Conformal prediction without logits. That a truncated candidate stage caps a downstream calibrated rule has precedent: cascaded conformal inference (Fisch et al., 2021) conditions its guarantee

on the truth surviving a cheap first stage, and distribution-free recommendation (Angelopoulos et al., 2023) turns on retrieval recall capping what is achievable. Both *choose* that stage; ours is imposed by the interface, which is what makes τ_k and the audit the contribution. Valid sets can also be built below L1, by repeated sampling (Su et al., 2024), calibrated stopping (Quach et al., 2024), filtering sampled claims (Mohri & Hashimoto, 2024; Cherian et al., 2024), or a handful of multiple-choice options (Ren et al., 2023; Kumar et al., 2023), the deployed workaround for this truncation (see Campos et al., 2024), and black-box uncertainty scores also sample (Lin et al., 2024). All are *constructions* at L3 or small K, and repeated sampling can estimate the hidden tail, so the hierarchy orders *per-call* access, not query complexity; our ceiling binds tail-blind rules on a single call's top-k response.

Censored and partial observation. Pessimistic score substitution for partially labeled data appears in Javanmardi et al. (2023); our achievability proof is the same move score-side, and we claim no novelty for it; conformalized survival analysis (Candès et al., 2023) calibrates through bounds on censored *outcomes*, as §8 does for censored scores. Semi-bandit online conformal prediction (Ge et al., 2025; Yang et al., 2026) censors the *outcome* and targets long-run coverage; we censor the *score* and keep the finite-sample guarantee. Shift-robust conformal methods (Tibshirani et al., 2019; Gibbs & Candès, 2021; Barber et al., 2023) reweight *scores*, which truncation withholds for censored labels. Nie et al. (2026) derive the identified set of a top-K-censored output distribution, which Proposition 3 transports to the conformal quantile; CPQ (Noorani et al., 2025) picks per-input sampling depth via Good–Turing (Good, 1953); our label space is enumerable, so under $\hat{p} = p$ the optimal policy is a closed-form threshold (Thm. 4). Scores that consume the full vector (Attar et al., 2026; Wang et al., 2026; Ding et al., 2026; Kotla & Kotla, 2025) are *inapplicable* through any interface we audit and motivate the question rather than compete on it; set-size-at-k measurements descend from OneC and size-stratified coverage (Vovk et al., 2016; Angelopoulos et al., 2021), and only their reading as a deployability constraint is ours.

3 Setup: the top-k oracle and the access hierarchy

Labels live in $\mathcal{Y} = \{1, \dots, K\}$ and f maps x to $\hat{p}(\cdot \mid x) \in \Delta^{K-1}$; write $\hat{p}_{(1)}(x) \geq \cdots \geq \hat{p}_{(K)}(x)$, $\mathrm{rank}(x, y) = 1 + \#\{y' : \hat{p}(y'|x) > \hat{p}(y|x)\}$ (ties, measure-zero offline, are broken by the oracle's order, so $|T_k(x)| = k$; live, the rank is the list position), $T_k(x) = \{y : \mathrm{rank}(x, y) \leq k\}$ and $P_k(x) = \sum_{r \leq k} \hat{p}_{(r)}(x)$.

Definition 1 (top-k oracle)**.** *The oracle $\mathcal{O}_k$ returns, for input x, the identities and probabilities of the k most likely labels: $\mathcal{O}_k(x) = \big((y_{(1)}, \dots, y_{(k)}),\ (\hat{p}_{(1)}(x), \dots, \hat{p}_{(k)}(x))\big)$.*

This is what a commercial API exposes, at the unit it scores (labels for a native classifier or next-token model, subword tokens for a chat endpoint; the token-to-label prefix map the audit uses is in §9.3(4) and Appendix B); the theory is stated at the label level. Data $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1})$ are exchangeable (i.i.d. where a result says so); the analyst observes $\mathcal{O}_k(X_i)$ and Y_i for $i \leq n$ plus $\mathcal{O}_k(X_{n+1})$, and must output $C(X_{n+1}) \subseteq \mathcal{Y}$ with $\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$; randomized scores may use auxiliary uniforms. Write $\mathrm{Acc}_k = \mathbb{P}(\mathrm{rank}(X, Y) \leq k)$ for top-k accuracy, $e_k = 1 - \mathrm{Acc}_k$, and $\mathrm{Recall}_k(y) = \mathbb{P}(\mathrm{rank}(X, Y) \leq k \mid Y = y)$.

Call a score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ *rank-monotone* if $\mathrm{rank}(x, y) \leq \mathrm{rank}(x, y')$ implies $s(x, y) \leq s(x, y')$, and *k-observable* if $s(x, y)$ is a measurable function of $(\mathcal{O}_k(x), y, U)$ whenever $\mathrm{rank}(x, y) \leq k$, U the auxiliary uniform drawn once per input.

Lemma 1. *LAC, randomized APS, RAPS and SAPS are rank-monotone and k-observable for all k.*

Rank-monotonicity makes every conformal set a *prefix* of the sorted label list, so with the truncated score of §5 "the API shows k labels" becomes "the prediction set is confined to those k" whenever $\hat{q} < \infty$. Interfaces come in four levels: **L0**, the full vector; **L1**, $\mathcal{O}_k(x)$; **L2**, ranks only (§6); **L3**, samples only (Noorani et al., 2025; Su et al., 2024); the order is by *per-call* disclosure (§2).

4 The ceiling: coverage is capped by top-k accuracy

Marginal validity under truncation is free (§5); the substantive question is *feasibility*: for which α does a nondegenerate valid predictor exist? The answer turns on the worth of the hidden tail.

Definition 2 (tail-concentration profile)**.** *For budget k and $b \geq 0$,*

$$\tau_k(b) \;=\; \mathbb{E}\Big[\textit{ sum of the } b \textit{ largest values of } \{\mathbb{P}(Y = y \mid \mathcal{O}_k(X)) : y \notin T_k(X)\}\Big],$$

extended to real b by linear interpolation (the fractional relaxation); τ_k is non-decreasing and concave in b.

$\tau_k(b)$ is the most coverage appending b hidden labels could buy given the response: always $\leq e_k$, $= b\,e_k/(K-k)$ under a conditionally uniform tail, and e_k when the response pins down the hidden truth. It is an oracle quantity; what calibration labels learn of it is a held-out lower bound (§4).

Theorem 1 (access bound)**.** *For i.i.d. data, let C be any set-valued rule measurable with respect to the calibration data D, $\mathcal{O}_k(X)$ and independent randomness (in particular any top-k rule), and suppose $|C(X) \setminus T_k(X)| \leq b$ almost surely. Then*

$$\mathbb{P}(Y \in C(X)) \;\leq\; \mathrm{Acc}_k \;+\; \tau_k(b).$$

The budget must bind per input, not on average. The hypothesis cannot be weakened to $\mathbb{E}|C(X) \setminus T_k(X)| \leq b$: a rule may concentrate its budget on the inputs whose model-implied hidden mass $1 - P_k(x)$, an observable, is largest. Appendix A gives a two-type tail-blind counterexample with mean budget 50 and coverage 1 against a bound of 0.95. Every use here has an almost-sure budget: $b = 0$ for rules confined to T_k, a fixed append-set size in §4.

Corollary 1 (tail-blind rules)**.** *Call C* tail-blind *if, conditionally on $\mathcal{O}_k(X)$ and D, the inclusion probabilities $\mathbb{P}(y \in C(X) \mid \mathcal{O}_k(X), D)$ of the $K-k$ hidden labels $y \notin T_k(X)$ are all equal. For tail-blind rules (or a conditionally uniform tail), again with $|C(X) \setminus T_k(X)| \leq b$ a.s.,*

$$\mathbb{P}(Y \in C(X)) \;\leq\; \mathrm{Acc}_k + \frac{b\,e_k}{K-k} \;\leq\; \mathrm{Acc}_k + \frac{b}{K-k}.$$

In the regime $b \ll K - k$ this reads $\alpha \geq e_k$: with k log-probabilities, the highest coverage a tail-blind rule can certify is the model's top-k accuracy. Equivalently, with the *minimum viable budget* $k_{\min}(\alpha) = \min\{k : e_k \leq \alpha\}$, certifying $1-\alpha$ requires being shown $k_{\min}(\alpha)$ entries.

Corollary 2 (class-conditional feasibility)**.** *Applying Theorem 1 conditionally on $\{Y = y\}$: a tail-blind rule whose append budget on the event $\{Y = y\}$ is at most b_y almost surely (so $b_y \leq b$), attaining class-conditional coverage $1-\alpha$ for class y, requires* $\mathrm{Recall}_k(y) \geq 1 - \alpha - b_y/(K-k)$.

As $\mathrm{Recall}_k(y) \ll \mathrm{Acc}_k$ for hard classes, class-conditional prediction dies long before marginal prediction (§9.1). The qualifier is essential: appending the audited class certifies it trivially.

Tail-blindness is not a technicality. The qualifier cannot be dropped: the bound $b/(K-k)$ does *not* hold for arbitrary rules, because the analyst observes Y_i for every calibration point, *including the censored ones*, so a rule can learn which hidden labels are frequently the truth. Every gain below is *held out*: mined by censored-label frequency on the calibration half, scored on the disjoint test half. On ImageNet at $k = 5$ the best learned append-set of size 20 buys 0.0044, or $3\times$ the tail-blind prediction, but only 6% of the deficit. On CLINC150 (Larson et al., 2019) a *single* label buys 0.0221 against 0.0003, because the out-of-scope intent is a genuine atom in the tail: mining it recovers 53% of the removed coverage, and live, appending that one calibration-identified label (always `oos`) to `gpt-4o-mini`'s output recovers **89%** of the deficit for one unit of set size (74–100% across six of seven; the seventh, whose window holds preamble rather than answers, recovers 16%; Table 10).

5 The floor: achievability, and when it collapses

Definition 3 (truncated score)**.** *For a rank-monotone k-observable s, define $s^{(k)}(x,y) = s(x,y)$ if* $\mathrm{rank}(x,y) \leq k$ *and $+\infty$ otherwise.*

Proposition 1 (marginal validity)**.** *$s^{(k)}(x,\cdot)$ is computable from $\mathcal{O}_k(x)$ alone. Split conformal with $s^{(k)}$ (threshold $\hat{q}$ the $\lceil (n+1)(1-\alpha) \rceil$-th smallest calibration score, $C(x) = \{y : s^{(k)}(x,y) \leq \hat{q}\}$) satisfies $\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1-\alpha$. Moreover $C(x) \subseteq T_k(x)$ whenever $\hat{q} < \infty$.*

This is elementary (pessimistic substitution, Javanmardi et al., 2023; rank truncation, Shi et al., 2024; Huang et al., 2024) and we claim it only as the matching half that makes the frontier $\alpha \geq e_k$ *tight*. The guarantee is marginal over D (Vovk, 2012; Bian & Barber, 2023), and truncation bites: given j censored calibration points and $\hat{q} < \infty$, coverage is exactly $(1 - e_k)\, m/(n - j + 1)$, $m = \lceil (n+1)(1-\alpha) \rceil$ (continuous scores; Appendix A). At ImageNet's $e_5 = 0.0712$, α=0.1, the sets that are usable cover 0.884 at n=50 and 0.893 at n=100 (within 10^{-3} of target from n≈500), while marginal coverage, 0.9113 at n=100, exceeds the upper band $1 - \alpha + \frac{1}{n+1}$ (which needs a.s. distinct scores): the degenerate sets pay for the shortfall. Live, the formula predicts partially degenerate cells to 0.002 (Table 9). The new object is exact finite-sample control of when the set is usable.

Theorem 2 (degeneracy law)**.** *Let s be finite on visible ranks (as all four scores of Lemma 1 are) and $N_{\text{cen}} = \#\{i \leq n : \text{rank}(X_i, Y_i) > k\}$. Then*

$$\hat{q} = +\infty \iff N_{\text{cen}} \geq \lfloor (n+1)\alpha \rfloor,$$

and for i.i.d. calibration data $N_{\text{cen}} \sim \text{Bin}(n, e_k)$, so $\mathbb{P}(\textit{degenerate}) = \mathbb{P}(\text{Bin}(n, e_k) \geq \lfloor (n+1)\alpha \rfloor)$ (hypergeometric for a random split of a fixed pool, the criterion unchanged).

Two consequences follow (Corollary 4, proved in the appendix). The transition at $e_k = \alpha$ is sharp: if $e_k < \alpha$ then $\mathbb{P}(\text{degenerate}) \leq \exp(-n\,\text{KL}(a_n \| e_k)) \to 0$ exponentially, with $a_n = \lfloor (n+1)\alpha \rfloor / n$ and once $a_n \geq e_k$ (Hoeffding, 1963), while if $e_k > \alpha$ it tends to 1 and at $e_k = \alpha$ to $1/2$. So $n \geq \log(1/\delta)/\text{KL}(a_n \| e_k)$, roughly $n \approx \text{KL}(\alpha \| e_k)^{-1} \log(1/\delta)$, calibration points suffice for degeneracy probability δ, and no n makes it vanish when $e_k \geq \alpha$.

This motivates the budget a practitioner needs: the *finite-sample budget* $k_{\min}(\alpha, n, \delta) = \min\{k : \mathbb{P}(\text{Bin}(n, e_k) \geq \lfloor (n+1)\alpha \rfloor) \leq \delta\}$. It is never smaller (strictly larger in every cell of Table 5), and at the frontier the gap does not close with n: at $e_k = \alpha = 0.1$ the degeneracy probability is 0.569 at n=50 and still 0.507 at n=5000, so there $k_{\min}(\alpha)$ is necessary and never sufficient. Inside the region Corollary 4 gives exponential decay, so finite n suffices, not small n: at ImageNet, α=0.1, k=5 ($e_5 = 0.0712$) it still degenerates with probability 0.283 at n=50, and on the LLM panel at n=2,000, δ=0.05 the finite-sample budget is 1.23–1.36× the population one at $\alpha \in \{0.1, 0.05\}$ and up to 1.97× at α=0.01 (Table 5). Theorem 4's constraint must likewise sit strictly below α.

The marginal guarantee puts the damage exactly where the model is weakest:

Proposition 2 (the censored slice is never covered)**.** *Any rule with $C(x) \subseteq T_k(x)$ has $\mathbb{P}(Y \in C(X) \mid \text{rank}(X, Y) > k) = 0$: the censored slice, the e_k-fraction of inputs the model finds hardest, contributes its entire mass to the miscoverage budget — e_k of the permitted α, all of it at $e_k = \alpha$.*

This is the truncated instance of a familiar fact, that distribution-free conditional coverage needs further assumptions (Barber et al., 2021), except that here the failing slice is *fixed in advance* by the interface (though not observable at test time), which makes it an audit: efficiency gains at fixed marginal coverage are bought by moving *all* permitted miscoverage onto that slice. Under $\hat{p} = p$ truncated LAC's large-n limit is size-optimal among rules confined to T_k (Theorem 5).

6 Rank-only, input-blind access forces fixed-size sets

One level down, the oracle returns the ordered labels and nothing else: the interface of recommender lists, N-best ASR, and every LLM endpoint exposing no logprobs. Three assumptions, the first doing the work: the rule is a measurable function of the L2 oracle output and D *only*, never of x itself; it is *identity-blind*: the set of *ranks* it selects is invariant under relabelings of the oracle output with D held fixed, so it operates on rank positions and D alone; and the model is *rank-calibrated*, $r \mapsto \mathbb{P}(\text{rank}(X, Y) = r)$ non-increasing. Call a rule *admissible* if it meets the first two, outputs a subset of the returned list (beyond it Cor. 1 governs), and covers at $1 - \alpha$. The first is strong: with identity-blindness it leaves no usable function of x, so "no adaptivity" is near-definitional and the content is the frontier itself.

Theorem 3 (rank-only access)**.** *Under these assumptions and i.i.d. data, every admissible rule is weakly dominated, at equal expected size and no less coverage, by one of the form $C(x) = T_R(x)$ with R a random variable independent of X. Consequently the Pareto frontier of achievable (coverage, size) pairs is the upper concave envelope of $\{(\text{Acc}_r, r) : r = 0, \ldots, k\}$; prefix rules randomizing R between adjacent integers attain every point of it, and split conformal on the truncated rank*

score $s^{(k)}$ with $s = \mathrm{rank}$ attains it as $n \to \infty$; and no adaptivity is possible*: for every admissible rule the law of the set size is the same at every input.*

The theorem is false without rank-calibration: with $K = 4$, $\hat{p} = (.5, .3, .15, .05)$ and true rank probabilities $(.2, .7, .07, .03)$, "predict the second-ranked label" attains coverage 0.70 at size 1, above the envelope. The assumption is weaker than calibration and checkable. Every one of our 33 profiles (29 full-scale, 4 minimal) shows some non-monotonicity in the first 50 ranks, but the largest violation is 1.3×10^{-3} and at most 3.05 SE against $\approx$1,600 adjacent-rank comparisons, which sampling noise alone would produce: we fail to reject the assumption rather than certify it. Prefix sets are always available, so the envelope is always *achievable*, and rank-calibration is what makes it unimprovable: the counterexample sits strictly *above* it. A richer rank-only rule, or one that also sees x, may need less size than the envelope, and truncated LAC is not the best L1 rule, so the comparison below bounds nothing. At matched coverage and $k \leq 20$ the envelope is a median $1.41\times$ the size of truncated LAC over the 81 (profile, α) cells where both reach $1 - \alpha$ (§B), range 1.02–2.15×; at the deepest cached $k = 1{,}000$ it is $1.50\times$ over 134 cells, below one in 8 vision cells where that LAC has full access and is far from the best L1 rule; live, LAC never returns the larger set on an endpoint that delivers L1, nor the rank score on the one that does not (Table 9).

7 Buying depth: the cost-optimal query policy

Depth can be bought (a higher tier, a self-hosted copy, logit-bias peeling, a cascade), so consider policies $k(\cdot)$ choosing a per-input depth at cost $\mathbb{E}[k(X)]$ entries. Validity is not at risk (a policy fixed before calibration leaves Proposition 1 intact); the question is cost.

Theorem 4 (optimal depth is a probability threshold)**.** *Among depth policies, the minimizer of $\mathbb{E}[k(X)]$ subject to the model-implied censoring constraint $\mathbb{E}[1 - P_{k(X)}(X)] \leq \alpha$ is*

$$k_\varepsilon(x) = \#\{y : \hat{p}(y \mid x) \geq \varepsilon\},$$

with ε calibrated to meet the constraint (boundary randomization included).

The Lagrangian separates across inputs: entry $r{+}1$ costs 1 and buys $\hat{p}_{(r+1)}(x)$ of mass, so reveal entries above $\varepsilon = 1/\mu$ (μ the multiplier). The optimum is a single *absolute* probability floor shared across inputs, the ϵ-sampling object of Hewitt et al. (2022), not the nucleus one (Holtzman et al., 2020), and not min-p (Nguyen et al., 2025), whose floor rescales by $\max_y \hat{p}(y|x)$ at each input:

Corollary 3. *Nucleus-style allocation $k_\rho(x) = \min\{k : P_k(x) \geq 1 - \rho\}$ with one global mass ρ (as Ravfogel et al. (2023) calibrate it for sampling, without entropy bins) is not optimal in general: it overspends on flat inputs, where the marginal probability of the next label is tiny while the cumulative target is far.*

If $\hat{p}$ is the true conditional law the model-implied constraint equals the realized censoring rate; otherwise they differ, by orders of magnitude on production endpoints (§9.3), so the realized-rate version is an empirical near-optimum, not a theorem (§9.1). A clairvoyant bound completes it: with labels known one pays $\mathrm{rank}(X_i, Y_i)$ on the $(1-\alpha)$-easiest points and nothing on the sacrificed slice, so $\mathbb{E}[k(X)] \geq \mathrm{OPT}(\alpha)$, the sum of the smallest $\lceil (1-\alpha)N \rceil$ ranks among N scored points divided by N.

8 Auditing your own information loss

A top-k user can bracket the full-access threshold $\hat{q}_{\mathrm{full}}$: a censored point has $\mathrm{rank}(X_i, Y_i) \geq k + 1$, giving two observable constraints, $\hat{p}(Y_i|X_i) \leq \hat{p}_{(k)}(X_i)$ and $\hat{p}(Y_i|X_i) \leq 1 - P_k(X_i)$.

Proposition 3 (identification bracket)**.** *Translating both through each score family yields per-example bounds $\ell_i \leq s(X_i, Y_i) \leq u_i$ computable from $\mathcal{O}_k$ (table in Appendix A); imputing censored scores at ℓ_i and at u_i and taking order statistics gives $\hat{q}_{\mathrm{lo}} \leq \hat{q}_{\mathrm{full}} \leq \hat{q}_{\mathrm{hi}}$ from top-k data alone.*

We claim no sharpness: with few hidden labels, sharing the leftover mass narrows the censored score further (Appendix A). The mass constraint does real work, binding for 51% of ImageNet calibration points at k=2; the bracket also collapses fast (LAC on ImageNet/ResNet-50, n=25,000: at α=0.1 its width is 0.123 at k=2, 0.017 at k=5, 0 from k=10), so a practitioner can certify from top-k data

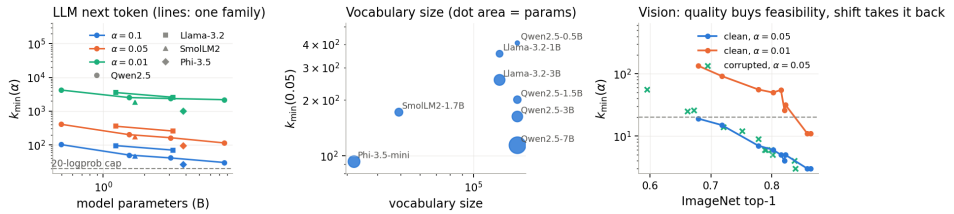

Figure 1: The feasibility frontier at scale: 400,000 full-vocabulary token positions and nine ImageNet checkpoints. **Left:** $k_{\min}(\alpha)$ vs. parameters, lines within one tokenizer family, the 20-logprob cap dashed. **Center:** $k_{\min}(0.05)$ against vocabulary size (dot area: parameters); descriptive only, vocabulary being collinear with family (§9.2). **Right:** on ImageNet quality buys feasibility ($k_{\min}(0.01)$ reaches 11) and corruption shift (crosses) takes it back.

that truncation costs nothing *at calibration* (the threshold is identified exactly; test-time sets still exclude the hidden labels, Prop. 2). The geometry is due to Nie et al. (2026).

9 Experiments

Offline, minimal: four cached label spaces: CLINC150 (K=151; TF-IDF + logistic regression), CIFAR-100, ImageNet/ResNet-50, GPT-2 next-token (K=50,257) (Larson et al., 2019; Krizhevsky, 2009; Russakovsky et al., 2015; Radford et al., 2019; Merity et al., 2017). **Offline, full scale:** eight open-weight LLMs (0.5B–7.6B) at full vocabulary, 50,000 positions each; nine timm checkpoints (Wightman, 2019) on ImageNet plus 12 (model, corruption) cells at severity 2 in the style of Hendrycks & Dietterich (2019), 270,000 scored images in all. **Live:** seven production endpoints, five tasks, 49,696 successful calls; everything below is post-hoc arithmetic on cached probabilities.

9.1 The frontier on four label spaces, and what validity costs

Feasibility and validity. $k_{\min}(\alpha)$ crosses the 20-logprob cap exactly where deployments want guarantees: ImageNet needs 26 at α=0.02 and 53 at 0.01, CIFAR-100 34 and 49, GPT-2 next-token 282 at α=0.1 and more than our k=4,096 horizon at 0.01; only CLINC150 stays inside the cap ($k_{\min}(0.01)$=19). Across 100 random half splits truncated LAC and APS cover within two Monte-Carlo standard errors of target in all 30 non-degenerate (dataset, score, α, k) cells (Table 4). Mean set size often *improves* (APS on ImageNet at α=0.1: 9.36 full-access vs. 4.43 at k=20), and Proposition 2 is the price tag: censored-slice coverage is identically zero in every cell, since truncation removes the inputs an adaptive score would spend on. Class-conditional feasibility dies first (Cor. 2): at k=20, α=0.05, where CIFAR-100 is marginally feasible, 17% of its classes are certifiably infeasible and 36% certifiably feasible (one-sided 95% Clopper–Pearson, 300 images per class).

Degeneracy and depth. Figure 2 (appendix) tracks the law through the phase transition. On realized censoring (on the model-implied constraint it wins by Theorem 4) the threshold policy beats a *flat* budget, one depth for every input, in 17 of 19 feasible (dataset, α) cells, ties in 1, loses in 1, saving up to 55% (tuned on a calibration half and scored held out: the same counts, up to 51%); nucleus loses to *flat* in 13 of its 17 feasible cells, and the clairvoyant bound sits below both.

9.2 Scale, vocabulary, and shift move the frontier; none of them rescue it

Figure 1 summarizes the full-vocabulary sweep; the full grid, with finite-sample budgets, is Appendix B.2. **(i) Scale helps; the panel cannot show saturation.** Within the one family that spans the range (Qwen2.5, one tokenizer) $k_{\min}(0.1)$ falls monotonically $102 \to 50 \to 41 \to 30$ from 0.5B to 7.6B (block-bootstrap 95% intervals $[94, 114]$ to $[26, 33]$), and at every α. Across families parameter count does not order the frontier (Llama-3.2-3B $[64, 76]$ against Qwen2.5-3B $[37, 46]$; Phi-3.5-mini at 3.8B reaches 26, below the 7.6B model): what looks like a stall past $\sim$3B is family as much as scale. At $\alpha = 0.01$ the best model needs 1,000, 50$\times$ the standard cap and 20$\times$ the deepest interface: in the range we can test, scaling does not close that gap. **(ii) The label space is**

Table 1: The seven-endpoint audit, two of five tasks; full grid in Table 6. Acc_k at the depth served, $\pm$ a 95% Wilson half-width; n is successful calls; *delivered* is the access level supplied at $T{=}0$ in default mode against the L1 all seven advertise († classification only; L1 with thinking off, §B.3). Last column: task-E mean label-set size $|S_k|$ at the served depth and coverage at a six-label budget (cov@6; §B); labels matched case-insensitively (§9.3(4)). **Task-A** Acc_k **is an upper bound** (a returned token counts when it prefixes the true continuation); the parenthesised value excludes tokens under two characters (finding 4). Lanes ran to launch size up to four refused calls (§B).

			A: next token		E: DBpedia ($K{=}219$)		
endpoint	cap	delivered	n	Acc_k ($\geq$2 char)	n	Acc_k	$\lvert S_k\rvert$ / cov@6
`gpt-4.1-mini`	20	L1	2000	0.949±0.010 (0.877)	1200	0.988±0.006	27.3 / 0.960
`gpt-4o-mini`	20	L1	2000	0.926±0.011 (0.825)	1200	0.978±0.009	18.7 / 0.928
`gpt-4.1-nano`	20	L1	2000	0.886±0.014 (0.762)	1200	0.944±0.013	21.7 / 0.797
`deepseek-chat`	20	**L2**	2000	0.949±0.010 (0.839)	1200	0.993±0.005	49.3 / 0.901
`qwen-plus`	5	L1	1999	0.851±0.016 (0.625)	1200	0.953±0.012	8.9 / 0.877
`qwen-turbo`	5	L1	1999	0.716±0.020 (0.524)	1200	0.909±0.016	9.3 / 0.853
`glm-4.7-maas`	50	**L3**†	800	0.990±0.007 (0.281)	1200	0.585±0.028	28.1 / 0.108

a second-order axis. Across the panel vocabulary is collinear with family (the two Acc_1-matched pairs of Table 5 differ by $+8\%$ and $+7\%$ in $k_{\min}(0.1)$, inside overlapping intervals). Holding the weights fixed and merging tokens whose strings agree up to case and whitespace shrinks V by 26–32% and $k_{\min}(0.1)$ by 7–14%. Neither base checkpoints nor fresh text rescue the cap: base models need up to $1.8\times$ fewer logprobs than their instruction-tuned twins, news published after every model's cutoff needs 1.2–$1.5\times$ more than the same outlet two years earlier (disjoint intervals, all 15 checkpoints), and no checkpoint reaches $k_{\min}(0.1) \leq 20$ on any corpus (minimum 25; §B.4). **(iii) Vision is feasible until shift.** The two strongest checkpoints reach $k_{\min}(0.01) = 11$, so the impossibility is not universal, but defocus blur takes the timm ResNet-50's $k_{\min}(0.05)$ from 6 to 56, across the cap, and an API user cannot see it happen: shift surfaces only as calibration-time degeneracy. Over 29 profiles the suite runs 6,380 single-split cells (§B), 2,924 degenerate (coverage one). On the remaining 3,456 the mean deviation is $+0.0013$ (sd 0.0064), worst cell -0.017, 4.4% of cells more than one point below target; conditioning on non-degeneracy selects calibration sets with few censored points (the shortfall of §5); counting degenerate cells at $+\alpha$ the unconditional mean is $+0.0220$. The depth policy of Theorem 4 beats a flat budget at equal realized censoring in 118 of the 134 (profile, α) cells where both are feasible, ties in 9 and loses in 7 (median saving where it wins, 42%; the losses, all vision at $\alpha \leq 0.02$, cost 1.13–$2.30\times$ the flat budget).

9.3 The live audit: what production endpoints actually deliver

We audited seven endpoints, all queried through one API account, on five tasks: next-token continuation on Wikipedia (A) and Python source (C) (Husain et al., 2019), and classification at $K = 151$ (CLINC150, B), 77 (Banking77, D; Casanueva et al., 2020) and 219 (DBpedia level-3, E): 49,696 calls in 35 cells. Four findings.

(1) Ceilings are tight and task-dependent (Table 1). The two best 20-cap Wikipedia cells coincide across providers (0.9490, 0.9485, ± 0.010), so the 20-logprob ceiling there lies between $\alpha \approx 0.05$ (every prefix match credited) and $\alpha \approx 0.12$ (single-character matches discarded, finding 4). Code is friendlier (0.974 raw, 0.886 at $\geq$2 characters). On 151-way classification six of seven fall below $\mathrm{Acc}_k = 0.83$, and on the answer-first ones because of one class: in-scope intents reach 0.92–1.00, the out-of-scope items (17–19%) 0–0.11; `deepseek-chat` reaches 0.88 on them, hence its 0.977 on an unpaired $n{=}300$ lane (the atom of §4). By Cor. 1 the ceiling binds every tail-blind procedure there.

(2) Reported probabilities cannot certify feasibility. Each response implies a censoring rate $1 - \sum_{r \leq k} \hat{p}_{(r)}$; the realized $1 - \mathrm{Acc}_k$ over its per-cell mean (§B) is 5–103$\times$ on next-token tasks and 57–7,856$\times$ on classification at answer-first endpoints (task D, `gpt-4.1-mini`: 4.1% realized, 0.0005% implied; infinite where the implied rate is 0). The implied rate forecasts the next token, the realized one the label, so the gap mixes token over-confidence, known after tuning (Kadavath et al., 2022; Tian et al., 2023), with the token–label mismatch of finding (4).

(3) Advertised access levels are not delivered access levels. All seven accept `top_logprobs`$= k$ and return well-formed L1 responses; two do not deliver it. `deepseek-chat` returns a top probability of *exactly* 1.0 on 100% of its 4,095 probability-bearing calls (of 4,100; five returned no window), all five tasks: every other probability is 0, so probability-weighted scores collapse to functions of the rank and the *delivered* level is **L2**. A temperature ablation identifies the mechanism: this endpoint returns *temperature-scaled* log-probabilities, one-hot at $T{=}0$ by construction (Table 7). The delivered level is a property of the *(interface, temperature)* pair that no schema documents: $T{=}0$, chosen for determinism, costs L1 on one route; the remedy is cheap, since temperature scaling preserves the ranking: request the window at $T{=}1$ and ignore the sampled token. On the ablation items truncated LAC at $\alpha{=}0.1$ then keeps 5.4 labels at 90.5%, against 43.1 at $T{=}0$ and below the rank score's 6.5 (Table 8): the probabilities are usable again, scaled or not, and $T \in \{1.7, 2\}$ buys nothing further. Even L1 routes return $p_{(1)}$ within 10^{-6} of 1 on 40–43% (`gpt-4.1-mini`) and 19–22% (`qwen-plus`) of calls at every T, a few tied float levels near score zero that LAC inherits. `glm-4.7-maas` is a reasoning endpoint queried in its default mode whose first token starts a chain of thought (`1` and `The` are 100% of task-E first positions), so 76% of its slots prefix no label. Its top is uninformative (28.1 of 219 labels survive at $\mathrm{Acc}_{50} = 0.585$): **L3** behaviour wherever the answer does not start at token one. Disabling thinking per request restores L1 ($\mathrm{Acc}_{50} = 0.995$; §B.3).

(4) A k-token window is not a k-label window. Endpoints score *tokens*, and every label beginning with a returned token stays alive: $S_k = \{l : \text{some top-}k \text{ token prefixes } l\}$. *Waste* (up to 95% of the window prefixes no label) and *collision* (one token keeps up to 3.8 alive) move the delivered budget to 0.18–2.5$\times$ the advertised depth. At matched delivered size (Table 1) the raw ranking reverses: `deepseek-chat`'s panel-best $\mathrm{Acc}_{20} = 0.993$ keeps 49.3 of 219 labels alive and falls to 0.901 at a six-label budget, below `gpt-4.1-mini` at 0.960 (disjoint intervals, identical items). The map is part of the interface, as surface-form competition predicts (Holtzman et al., 2021; Zhao et al., 2021): the prompt asks for lowercase labels, DBpedia's are CamelCase, and an exact-case map would censor exactly those answers (`gpt-4.1-mini`'s task-E Acc_1: 0.006, not 0.870). Numbering the task-B labels raises raw Acc_{20} by up to $+0.044$ but loses at a matched five-label budget (§B.1): the ceiling is joint in interface and protocol.

Running the procedure on the live calls. Split conformal on the task-E calls (truncated LAC and rank scores, sizes in *labels*; Table 9) degenerates where the ceiling binds: at $\alpha{=}0.05$ two endpoints return all 219 labels on every split and two more, with e_k within 0.006 of α, on 82% and 37% of splits. On `deepseek-chat`, whose constant probabilities make LAC two-valued, it keeps 49.4 labels at $\alpha{=}0.05$ where the rank score gives 7.9 at 95.9%, 6.3$\times$ fewer. **Against sampling (L3).** On 1,000 of these items, splits shared with the window, 20 answers at $T{=}1$ scored as in Su et al. (2024) resolve what the window cannot (about one label at $\alpha{=}0.2$ where feasible, against 2.3–5.3), but the sampled pool is the harsher truncation (`gpt-4.1-mini`: $\mathrm{Acc}^{(20)} = 0.910$, $\mathrm{Acc}_{20} = 0.988$), so at $\alpha{=}0.05$ it degenerates on every split on three routes and on 58% of splits on `deepseek-chat`, while the window certifies three of the four (§B.3).

10 Conclusion

We asked how much distribution-free coverage survives when an interface returns only its top k scores. For rules built on the response alone the answer is top-k accuracy, the truncated procedure reaches it marginally, and a binomial law says when it collapses. The production endpoints we audited put that cap where deployments want guarantees, and their returned probabilities cannot locate it; labeled calls can.

What we claim, and what we do not. Truncation keeps marginal coverage and, offline, often shrinks mean sets; the cost is conditional, on the censored slice (Prop. 2) and near the frontier on the calibration set (§5). The ceiling needs an almost-sure append budget and binds rules built on the interface alone (an analyst's own model of x can escape it); reading it as $\approx\mathrm{Acc}_k$ needs tail-blindness (§4). How hard it binds varies: the best classification routes clear $\alpha{=}0.05$ on DBpedia, open-vocabulary outputs need 26–102 logprobs at $\alpha{=}0.1$. The L1-vs-L2 gap compares procedures, the frontier is marginal and i.i.d. (shift is out of scope), and live results are the routes as served to one account.

ETHICS STATEMENT

This work audits the behavior of publicly purchasable APIs using public-benchmark inputs; no user data, personal data, or adversarial extraction beyond documented `logprobs` parameters was involved. The audit findings (degenerate probability channels, preamble-burning windows) are reported at the level of served interfaces, not training data or model internals, and are reproducible for a modest API spend on the endpoints as served in August 2026; the shipped records are the durable artifact. Logit-bias peeling is cited as prior art, not performed. We see no plausible harm vector beyond embarrassment for the routes as served, against which we weigh the safety benefit of practitioners not over-trusting certified coverage that their interface cannot support.

REPRODUCIBILITY STATEMENT

Every result is proved in Appendix A with assumptions stated inline (Proposition 2 is immediate from $C \subseteq T_k$). The supplementary material contains: the complete code (collection, analysis, and figure generation), the 28-endpoint census probes, the raw per-call API records (49,696 audit calls, the 4,000-call protocol ablation of §B.1 and the 3,598-call temperature ablation of Table 7, and the September 2026 follow-up records of §B.3, all as shipped JSONL), the subsampled CIFAR-100 and GPT-2 score caches and the rank histograms behind the LLM panel (the full $\sim$1.1 GB caches regenerate from the shipped scripts, and the externally hosted ImageNet logits are downloaded by the shipped fetch script rather than redistributed), per-cell result JSON for all 6,380 suite cells, and run notes documenting every operational failure and correction (the item-pairing structure of the task-B and task-D lanes launched at different n, recovered from stored truth prefixes and stated in §B, and the single-character restatement of §9.3). Failed calls retain their exception class and HTTP status; the provider's free-text error bodies, which carry only routing and billing detail, are redacted. The shipped analysis scripts regenerate every table and figure in this paper from the shipped records and summaries; re-scoring the eight open-weight checkpoints and nine vision models needs a GPU and the public weights.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We did not use generative AI tools for any of the other tasks that require disclosure: in particular, we did not use them to generate synthetic data, to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support data analysis or the interpretation of results. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios N Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Anastasios N Angelopoulos, Karl Krauth, Stephen Bates, Yixin Wang, and Michael I Jordan. Recommendation systems with distribution-free reliability guarantees. In *Symposium on Conformal and Probabilistic Prediction with Applications*, 2023.

Navid Akhavan Attar, Hesam Asadollahzadeh, Ling Luo, and Uwe Aickelin. Softmax is not enough (for adaptive conformal classification). In *International Conference on Learning Representations*, 2026.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference*, 10(2):455–482, 2021.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Michael Bian and Rina Foygel Barber. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066, 2023.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 2024.

Emmanuel Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45, 2023.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Eric Wallace, David Rolnick, and Florian Tramèr. Stealing part of a production language model. In *International Conference on Machine Learning*, 2024.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Workshop on Natural Language Processing for Conversational AI*, 2020.

John J Cherian, Isaac Gibbs, and Emmanuel J Candès. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, and Titouan Lorieul. Set-valued classification — overview via a unified framework. *arXiv preprint arXiv:2102.12318*, 2021.

Christophe Denis and Mohamed Hebiri. Confidence sets with expected sizes for multiclass classification. *Journal of Machine Learning Research*, 18(102):1–28, 2017.

Tiffany Ding, Anastasios N Angelopoulos, Stephen Bates, Michael I Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tiffany Ding, Jean-Baptiste Fermanian, and Joseph Salmon. Conformal prediction for long-tailed classification. In *International Conference on Learning Representations*, 2026.

Matthew Finlayson, Xiang Ren, and Swabha Swayamdipta. Logits of API-protected LLMs leak proprietary information. In *Conference on Language Modeling*, 2024.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Efficient conformal prediction via cascaded inference with expanded admission. In *International Conference on Learning Representations*, 2021.

Haosen Ge, Hamsa Bastani, and Osbert Bastani. Stochastic online conformal prediction with semi-bandit feedback. In *International Conference on Machine Learning*, 2025.

Isaac Gibbs and Emmanuel J Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

John Hewitt, Christopher D Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, 2022.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. In *International Conference on Machine Learning*, 2024.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. CodeSearchNet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*, 2019.

Alireza Javanmardi, Yusuf Sale, Paul Hofman, and Eyke Hüllermeier. Conformal prediction with partially labeled data. In *Conformal and Probabilistic Prediction with Applications*, pp. 251–266, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Yoshith Roy Kotla and Varshith Roy Kotla. Conformal prediction sets for next-token prediction in large language models: Balancing coverage guarantees with set efficiency. *arXiv preprint arXiv:2512.22682*, 2025.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Empirical Methods in Natural Language Processing*, 2019.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024.

Rui Luo and Zhixin Zhou. Trustworthy classification through rank-based conformal prediction sets. *arXiv preprint arXiv:2407.04407*, 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *International Conference on Machine Learning*, 2024.

John X. Morris, Wenting Zhao, Justin T. Chiu, Vitaly Shmatikov, and Alexander M. Rush. Language model inversion. In *International Conference on Learning Representations*, 2024.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *International Conference on Learning Representations*, 2025.

Wenhua Nie, ZiCheng Zhu, Jianan Wu, Binhan Luo, Haoran Zheng, and Jyh-Shing Roger Jang. Identified-set geometry of distributional model extraction under top-K censored API access. *arXiv preprint arXiv:2605.10407*, 2026.

Sima Noorani, Shayan Kiyani, George Pappas, and Hamed Hassani. Conformal prediction beyond the seen: A missing mass perspective for uncertainty quantification in generative models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2506.05497.

Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alexander Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pp. 345–356, 2002.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 2019.

Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. Conformal nucleus sampling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 27–34, 2023.

Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning*, 2023.

Yaniv Romano, Matteo Sesia, and Emmanuel J Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Yuanjie Shi, Subhankar Ghosh, Taha Belkhouja, Janardhan Rao Doppa, and Yan Yan. Conformal prediction for class-wise coverage via augmented label rank calibration. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2406.06818.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. API is enough: Conformal prediction for large language models without logit-access. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 979–995, 2024.

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

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel J Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for conformal prediction. In *Symposium on Conformal and Probabilistic Prediction with Applications*, 2016.

Tao Wang, Yan Sun, and Edgar Dobriban. Singleton-optimized conformal prediction. In *International Conference on Learning Representations*, 2026.

Ross Wightman. PyTorch image models. `https://github.com/huggingface/pytorch-image-models`, 2019.

Junyoung Yang, Kyungmin Kim, and Sangdon Park. Online conformal prediction with adversarial semi-bandit feedback via regret minimization. In *International Conference on Learning Representations*, 2026.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, 2021.

A Proofs

A.1 Proof of Lemma 1

Write $q_r = \hat{p}_{(r)}(x)$ and $c_r = \sum_{j<r} q_j$. LAC: $s = 1 - q_r$ is non-decreasing in r since q_r is non-increasing; q_r is in $\mathcal{O}_k(x)$ for $r \le k$. Randomized APS: $s = c_r + Uq_r$ with $U \in [0,1]$; $s_{r+1} - s_r = (1-U)q_r + Uq_{r+1} \ge 0$; computing s at rank $r \le k$ needs $q_1, \dots, q_r$, all observable. RAPS adds $\lambda(r - k_{\text{reg}})_+$, non-decreasing in r and a function of r only. SAPS: $s = Uq_1$ at $r = 1$ and $q_1 + (r - 2 + U)\lambda$ for $r \ge 2$; the second branch increases in r and at $r = 2$ equals $q_1 + U\lambda \ge Uq_1$; it needs only q_1 and r. □

A.2 Proof of Theorem 1

Work conditionally on D throughout and write $O = \mathcal{O}_k(X)$; $T_k(X)$ is a function of O (the returned identities), written $T_k(O)$. Split on $A = \{Y \in T_k(O)\}$: $\mathbb{P}(Y \in C \mid D) \le \mathrm{Acc}_k + \mathbb{P}(Y \in C, A^c \mid D)$. Write $\pi_y(o) = \mathbb{P}(y \in C \mid O = o, D)$ and $\nu_o(y) = \mathbb{P}(Y = y \mid O = o)$; because the test pair is independent of D (i.i.d. sampling), Y given (O, D) has law ν_O, and C depends on the test pair only through O, its own randomization being independent of Y, so

$$\mathbb{P}(Y \in C, A^c \mid D) = \mathbb{E}_O\Big[\textstyle\sum_{y \notin T_k(O)} \pi_y(O)\, \nu_O(y)\Big].$$

(Under exchangeability alone D can carry information about Y given O, and the factorization fails; so it does for a rule that also reads X, whose inclusion probabilities can correlate with Y given O: see the example after the proof of Corollary 1.) For each o, the inner sum is a linear functional of π over $\{0 \le \pi_y \le 1,\ \sum_{y \notin T_k(o)} \pi_y = b(o)\}$ with $b(o) = \mathbb{E}[|C \setminus T_k(o)| \mid O = o, D] \le b$, maximized by putting mass 1 on the $b(o)$ largest values of $\nu_o(y)$ over $y \notin T_k(o)$ (fractionally on the last, if $b(o) \notin \mathbb{Z}$). Write $g(o, \beta)$ for that maximum: it is the running sum of the β largest hidden conditional probabilities given o, hence non-decreasing in β, and $\mathbb{E}_O[g(O, \beta)] = \tau_k(\beta)$ for constant β. Since $b(o) \le b$ almost surely, $g(o, b(o)) \le g(o, b)$ pointwise, and taking expectations gives $\mathbb{P}(Y \in C, A^c \mid D) \le \mathbb{E}_O[g(O, b)] = \tau_k(b)$; the bound holds for every D, hence unconditionally.

Monotonicity, not concavity, is what the argument needs. Bounding only $\mathbb{E}[b(O)]$ would require $\mathbb{E}_O[g(O, b(O))] \le \mathbb{E}_O[g(O, \mathbb{E}b(O))]$, which is not an instance of Jensen's inequality because g depends on o as well as on the budget. When O determines X and Y is a function of X, $\tau_k(b) = e_k$ for every $b \ge 1$ and the bound is vacuous; it is attained by the rule that appends the hidden truth, so it is tight, but only a rule that learns that map from O reaches it; Corollary 1 is the statement with content for rules that do not single out hidden labels.

Counterexample under a mean budget (the remark in §4). Let $K = 101$, $k = 1$, and let X take two equally likely values: at x_1, $\hat{p} = (0.8, 0.002, \dots, 0.002)$; at x_2, $\hat{p} = (1, 0, \dots, 0)$, with $\hat{p}$ the true conditional law. Then $\mathrm{Acc}_1 = 0.9$, $e_1 = 0.1$, and the hidden tail is conditionally uniform, so $\tau_1(b) = b\, e_1/(K - k) = 0.001\, b$ exactly. Consider $C(x_1) = \mathcal{Y}$ and $C(x_2) = \{y_{(1)}\}$: it is

measurable with respect to $\mathcal{O}_k(X)$ (the two types differ in $\hat{p}_{(1)}$, which is returned) and tail-blind (hidden inclusion probabilities are identically 1 at x_1 and identically 0 at x_2). Its mean budget is $\mathbb{E}|C(X) \setminus T_1(X)| = 50$, so the mean-budget statement would assert coverage $\leq \mathrm{Acc}_1 + \tau_1(50) = 0.95$; its actual coverage is 1. The almost-sure budget for this rule is 100, and $\mathrm{Acc}_1 + \tau_1(100) = 1$, which the true coverage attains with equality. □

A.3 Proof of Corollary 1

Tail-blindness makes $\pi_y(o)$ constant over hidden labels $y \notin T_k(o)$, so the inner sum equals $\frac{b(o)}{K-k} \sum_{y \notin T_k(o)} \nu_o(y) \leq \frac{b}{K-k} \sum_{y \notin T_k(o)} \nu_o(y)$ by the almost-sure budget, and its expectation is at most $\frac{b}{K-k} e_k$; the second inequality uses $e_k \leq 1$. Under a conditionally uniform tail (given O), each hidden label carries mass $\frac{1}{K-k} \sum_{y \notin T_k(o)} \nu_o(y)$ and $\tau_k(b) = b\, e_k/(K-k)$ directly. □

Why the rule may not read x. Let $K = 3$, $k = 1$, and two equiprobable inputs a, a' return the same window (top label 1, $\hat{p} = (0.5, 0.25, 0.25)$), with $Y \mid a$ uniform on $\{1, 2\}$ and $Y \mid a'$ uniform on $\{1, 3\}$. The rule $C(a) = \{1, 2\}$, $C(a') = \{1, 3\}$ appends one hidden label, each with probability $1/2$ given the response, so it is tail-blind in the sense of Corollary 1 with $b = 1$; it covers with probability 1, above $\mathrm{Acc}_1 + b\, e_1/(K-k) = 0.75$. It is excluded because it reads x beyond $\mathcal{O}_k(x)$, the interface's whole content; the same restriction carries Corollary 2.

A.4 Proof of Corollary 2

Condition the argument of Theorem 1 on $\{Y = y\}$: coverage of class y splits into the visible part, at most $\mathrm{Recall}_k(y)$, and the hidden part, at most $b_y/(K-k)$, since a tail-blind rule measurable as in Theorem 1 includes a hidden y with probability $\pi_y(O) = b(O)/(K-k)$ whatever the value of Y, and $b(O) \leq b_y$ on $\{Y = y\}$. Requiring $1 - \alpha \leq \mathrm{Recall}_k(y) + b_y/(K-k)$ rearranges to the statement. □

A.5 Proof of Proposition 1

For $y \in T_k(x)$, s is k-observable by assumption; for $y \notin T_k(x)$, membership is decidable from $\mathcal{O}_k(x)$ (the identities of the top k are returned) and the value is the constant $+\infty$. Hence $s^{(k)}$ is a measurable function of the observed data, and the standard split-conformal lower guarantee (Vovk et al., 2005; Lei et al., 2018) applies verbatim: coverage at least $1 - \alpha$. If $\hat{q} < \infty$, then every $y \notin T_k(x)$ has score $+\infty > \hat{q}$, so $C(x) \subseteq T_k(x)$.

The matching upper band does *not* transfer. Its proof needs the $n + 1$ scores to be almost surely distinct, and $s^{(k)}$ has an atom of mass e_k at $+\infty$ by construction; distinctness of the finite scores does not remove it. Coverage is at least $\mathbb{P}(\hat{q} = +\infty)$, since on that event $C(x) = \mathcal{Y}$. With $e_k = 0.3$, $n = 50$, $\alpha = 0.1$, Theorem 2 gives $\mathbb{P}(\hat{q} = +\infty) = \mathbb{P}(\mathrm{Bin}(50, 0.3) \geq 5) = 0.9998$, so coverage is 0.9999 against a putative band of 0.9196; at ImageNet's $e_5 = 0.071$ with $\alpha = 0.1$, $n = 100$, simulation gives 0.9113 against 0.9099. The band is recovered when $e_k = 0$.

Coverage given the calibration set. Condition on $N_{\mathrm{cen}} = j < \lfloor (n+1)\alpha \rfloor$ (so $\hat{q} < \infty$). The $n - j$ visible calibration scores and a visible test score are i.i.d. given visibility, so a visible test point is covered with probability exactly $m/(n - j + 1)$ (continuous scores), and a censored one, which occurs with probability e_k independently of D, never is: coverage is $(1 - e_k)\, m/(n - j + 1)$, increasing in j. Averaging over the non-degenerate j with $\mathrm{Bin}(n, e_k)$ weights gives coverage on the event that the set is usable: 0.884, 0.893, 0.897 at $n = 50, 100, 200$ for $e_k = 0.0712$, $\alpha = 0.1$, and within 10^{-3} of $1 - \alpha$ from $n \approx 500$; adding the degenerate event at coverage one returns the marginal value. For a random split of a fixed pool, replace e_k by the censored share of the test half and the weights by hypergeometric ones: this predicts 0.930, 0.943 and 0.895 for the three partially degenerate cells of Table 9, against 0.931, 0.945 and 0.896 realized. Neither the band nor $1 - \alpha$ holds conditionally on non-degeneracy. □

A.6 Proof of Theorem 2

Let $m = \lceil (n + 1)(1 - \alpha) \rceil$; $\hat{q}$ is the m-th order statistic of n scores of which exactly N_{cen} equal $+\infty$. It is infinite iff $m > n - N_{\mathrm{cen}}$, i.e. $N_{\mathrm{cen}} > n - m$. Since $m = (n + 1) - \lfloor (n + 1)\alpha \rfloor$, we

get $n - m = \lfloor (n+1)\alpha \rfloor - 1$, so $\hat{q} = +\infty \iff N_{\text{cen}} \geq \lfloor (n+1)\alpha \rfloor$. For i.i.d. data, censoring indicators are i.i.d. Bernoulli(e_k). □

A.7 Corollary 4 (phase transition) and its proof

Corollary 4 (phase transition and calibration-size requirement)**.** *With $a_n = \lfloor (n+1)\alpha \rfloor / n \to \alpha$: if $e_k < \alpha$ then, for every n with $a_n \geq e_k$ (guaranteed once $n \geq (1-\alpha)/(\alpha - e_k)$), $\mathbb{P}(\text{degenerate}) \leq \exp(-n \cdot \mathrm{KL}(a_n \| e_k)) \to 0$ exponentially; if $e_k > \alpha$ then $\mathbb{P}(\text{degenerate}) \to 1$; if $e_k = \alpha$ then $\mathbb{P}(\text{degenerate}) \to 1/2$. Hence $n \geq \log(1/\delta)/\mathrm{KL}(a_n \| e_k)$ suffices for degeneracy probability δ ($\approx \mathrm{KL}(\alpha \| e_k)^{-1} \log(1/\delta)$ since $a_n \to \alpha$), and no n makes it vanish when $e_k \geq \alpha$.*

Apply the Chernoff–Hoeffding bound in relative-entropy form (Hoeffding, 1963), $\mathbb{P}(\mathrm{Bin}(n,p) \geq na) \leq \exp(-n\,\mathrm{KL}(a\|p))$ for $a \geq p$, to $\mathbb{P}(\mathrm{Bin}(n, e_k) \geq n a_n)$; the precondition $a_n \geq e_k$ is where the small-n restriction comes from ($a_n \geq \alpha - (1-\alpha)/n$), and without it the bound is false (at $a_n = 0$ the left side is 1). For $e_k > \alpha$, the law of large numbers puts $N_{\text{cen}}/n \to e_k$ above the threshold eventually; at $e_k = \alpha$ the threshold sits within 1 of the mean, so the central limit theorem gives $1/2$. Solving $\exp(-n\,\mathrm{KL}(a_n \| e_k)) \leq \delta$ gives a sufficient sample size; a_n is below α for most n, so replacing it by α is an approximation, not a bound, and the exact binomial criterion of Theorem 2 is what a practitioner should solve. We do not claim the count necessary. □

A.8 Size-optimality of truncated LAC

Theorem 5 (truncated LAC is size-optimal under the idealized model)**.** *Suppose the data are i.i.d., $\hat{p}(\cdot|x)$ is the true conditional law of Y given $X = x$, and $e_k \leq \alpha$, so that some rule confined to T_k covers at $1 - \alpha$ (Theorem 1 with $b = 0$ shows none does otherwise). Among all rules with $C(x) \subseteq T_k(x)$ and $\mathbb{P}(Y \in C(X)) \geq 1 - \alpha$, expected size $\mathbb{E}|C(X)|$ is minimized by $C^*(x) = \{y \in T_k(x) : \hat{p}(y|x) \geq t\}$ with t (and boundary randomization) chosen so coverage is exactly $1 - \alpha$.*

Proof. The problem is a fractional knapsack over pairs (x, y) with $y \in T_k(x)$: including y at x contributes $\hat{p}(y|x)\, d\mathbb{P}(x)$ to coverage at cost $d\mathbb{P}(x)$ in size, so the value-to-cost ratio is $\hat{p}(y|x)$, and greedy inclusion in decreasing ratio order (a global threshold on $\hat{p}(y|x)$ intersected with $T_k(x)$, randomized at the boundary) is optimal for the continuous relaxation and implementable exactly. This is Sadinle et al. (2019) restricted to the truncated feasible set. □

On real data with an imperfect model the optimality can fail, and does in 3 of our cells; we report it as the idealized benchmark.

A.9 Proof of Theorem 3

By hypothesis the rule is a function of the L2 oracle output and D alone, and any injective k-tuple of labels is carried to any other by some relabeling, so identity-blindness makes the *set of ranks* selected a function of D (and auxiliary randomness) alone: its conditional law given D, call it $\mathcal{S}$, is the same for every x, which is the no-adaptivity clause. (If the rule may also read x, or may match the returned identities against D's label counts, this step fails and R may vary with the input.) Under i.i.d. sampling selecting rank r yields coverage $\mathbb{P}(\mathrm{rank}(X, Y) = r)$. If $\mathcal{S}$ is not a prefix, with $r \in \mathcal{S}$, $r' < r$ and $r' \notin \mathcal{S}$, swapping r for r' preserves size and changes coverage by $\mathbb{P}(\mathrm{rank} = r') - \mathbb{P}(\mathrm{rank} = r) \geq 0$, where the inequality is exactly rank-calibration. Hence every admissible rule is weakly dominated by a prefix rule $\mathcal{S} = \{1, \ldots, R\}$ with $R \perp X$, coverage $\mathbb{E}[\mathrm{Acc}_R]$, size $\mathbb{E}[R]$; a non-prefix rule remains admissible whenever its coverage clears $1 - \alpha$ (a rule selecting ranks $\{1, 3\}$, say) but attains no more coverage at its size than the prefix rule of that size, so the Pareto frontier, not the set of all achievable pairs, is the upper concave envelope of $\{(\mathrm{Acc}_r, r)\}$. A prefix rule randomizing R between adjacent integers attains any point of the envelope. Split conformal on the truncated rank score selects $R = \hat{q}$, the $\lceil (n+1)(1-\alpha) \rceil$-th order statistic of calibration ranks; its operating point $(\mathbb{E}[\mathrm{Acc}_{\hat{q}}], \mathbb{E}[\hat{q}])$ is a mixture over the law of $\hat{q}$ and lies on the envelope only when that law is supported on two adjacent integers, which holds as $n \to \infty$ ($\hat{q}$ concentrates on the population $(1-\alpha)$-quantile of the rank). Without rank-calibration the swap step fails and the counterexample in the main text applies. □

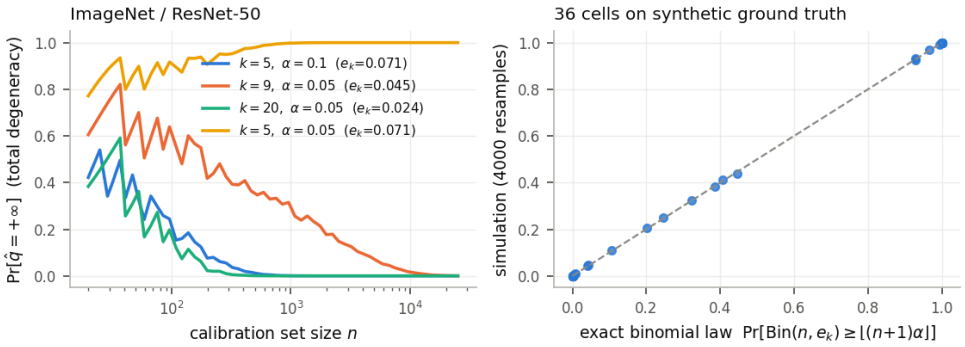

Figure 2: Theorem 2, two ways. **Left:** the law on ImageNet/ResNet-50 at four (k, α) cells. The transition at $e_k = \alpha$ is sharp (the three cells with $e_k < \alpha$ decay to zero, the one with $e_k > \alpha$ saturates at one), and degeneracy stays substantial at practical n inside the feasible region. **Right:** predicted against realized degeneracy over 36 synthetic cells, 4,000 resampled calibration sets each. The if-and-only-if criterion itself matched the realized $\hat{q}$ in 10,800 of 10,800 further resampled sets (the same 36 cells, 300 each). Synthetic cells draw i.i.d. Bernoulli censoring; a random split of a fixed pool follows the hypergeometric form (Table 9).

A.10 Proof of Theorem 4 and Corollary 3

The Lagrangian $\mathbb{E}[k(X)] + \mu(\mathbb{E}[1 - P_{k(X)}(X)] - \alpha)$ separates across x. At fixed x, moving from depth r to $r+1$ costs 1 and reduces uncovered mass by $\hat{p}_{(r+1)}(x)$: profitable iff $\mu\,\hat{p}_{(r+1)}(x) > 1$. Since $\hat{p}_{(r)}(x)$ is non-increasing in r, the per-input optimum is $\#\{y : \hat{p}(y|x) \geq 1/\mu\}$; sweeping μ traces the constraint, with randomization at the boundary atom meeting it exactly, and for any feasible policy k, $\mathbb{E}[k] \geq \mathbb{E}[k] + \mu(\mathbb{E}[1 - P_k] - \alpha) \geq \min_{k'} L(k', \mu) = \mathbb{E}[k_\varepsilon]$ when the constraint binds. For the corollary, a two-type instance: x_1 with $\hat{p} = (0.9, 0.1)$ and x_2 with $\hat{p} = (0.3, 0.05 \times 14)$, equiprobable, $\alpha = 0.1$. Nucleus with mass 0.9 spends depths $(1, 13)$ at cost 7.0; mixing $\varepsilon = 0.05$ (depths $(2, 15)$, censoring 0) with $\varepsilon = 0.1$ (depths $(2, 1)$, censoring 0.35) at weight 0.714 meets 0.1 at cost 6.5. Equality cases exist (inputs sharing one tail profile), hence "in general"; the 13-of-17 count is measured in §9.1. □

A.11 Proof of Proposition 3, with the per-score bounds

A censored point has $\mathrm{rank}(X_i, Y_i) \geq k+1$, hence $\hat{p}(Y_i|X_i) \leq \hat{p}_{(k)}(X_i)$ (it is beaten by k labels) and $\hat{p}(Y_i|X_i) \leq 1 - P_k(X_i)$ (hidden labels share the leftover mass). Translating through each family gives observable bounds $\ell_i \leq s(X_i, Y_i) \leq u_i$, where $U_i \in [0, 1]$ is the analyst's own auxiliary uniform for point i (drawn by the analyst, hence known):

score	ℓ_i	u_i
LAC	$\max(1 - \hat{p}_{(k)},\ P_k)$	1
APS	P_k	1
RAPS	$P_k + \lambda(k{+}1 - k_{\mathrm{reg}})_+$	$1 + \lambda(K - k_{\mathrm{reg}})_+$
SAPS	$\hat{p}_{(1)} + (k{-}1{+}U_i)\lambda$	$\hat{p}_{(1)} + (K{-}2{+}U_i)\lambda$

Imputing censored scores at ℓ_i (resp. u_i) can only lower (resp. raise) the m-th order statistic, since order statistics are monotone in each argument; hence $\hat{q}_{\mathrm{lo}} \leq \hat{q}_{\mathrm{full}} \leq \hat{q}_{\mathrm{hi}}$. The bounds use two constraints only and are not sharp in general: the $K - k$ hidden labels share $1 - P_k$ with each at most $\hat{p}_{(k)}$, so $\hat{p}(Y_i|X_i) \geq \max(0,\ 1 - P_k - (K - k - 1)\,\hat{p}_{(k)})$, which binds when $K - k$ is small (K=3, k=1, $\hat{p}_{(1)} = 0.4$ confines a censored LAC score to $[0.6, 0.8]$, not $[0.6, 1]$); for $K \gg k$ that constraint is inactive. □

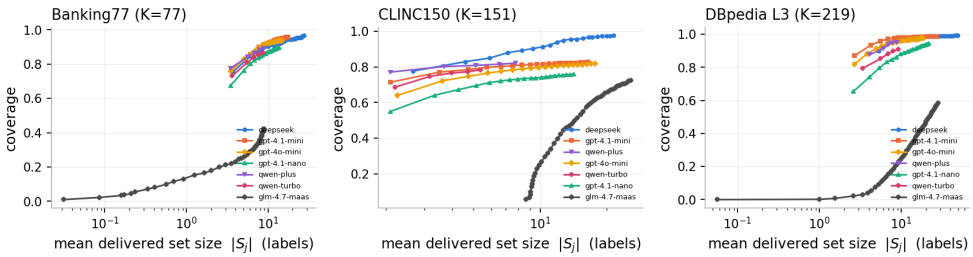

Figure 3: The label-space frontier behind a chat API, the curves behind Table 1's last column: coverage of the delivered label set S_j against its mean size as query depth j walks $1..k$, per endpoint and task. Equal depth is not equal difficulty: endpoints whose windows are full of short tokens buy coverage with set size, so comparisons are honest only at matched delivered size, where the raw ranking reverses (`deepseek` vs. `gpt-4.1-mini` on DBpedia, right panel).

B Experimental details and additional results

Terms. A *profile* is one cached score matrix, one model on one dataset (29 full-scale: 8 LLM, 9 clean and 12 shifted vision; 4 minimal). A suite *cell* is one run of split conformal at one (profile, score, α, k) on one seeded half split. A *lane* is the calls of one endpoint on one task; a *route* is one endpoint identifier. Two procedures are compared at *matched coverage* when both are set to the same nominal $1-\alpha$: conformal ones by calibration, the rank envelope exactly, by randomizing between adjacent prefixes.

Conventions. Ranks use the optimistic tie convention $\mathrm{rank}(x, y) = 1 + \#\{y' : \hat{p}(y') > \hat{p}(y)\}$; on ImageNet the pessimistic convention changes Acc_1 from 0.76182 to 0.76142. All conformal indices use $\lceil (n+1)(1-\alpha) \rceil$ with a 10^{-9} floating-point guard. Token positions within a document are autocorrelated, so a sequence of positions is not exchangeable. Feasibility quantities (Acc_k, $k_{\min}$) are marginal statistics, whose uncertainty we report with a block bootstrap over the 512-position windows the full-scale runs score (consecutive positions inside a window, disjoint windows); validity checks split a fixed corpus at random, so the split is exchangeable by construction and the check certifies the finite-population guarantee, not coverage on a future document, which is a non-exchangeable question (Barber et al., 2023) we do not address.

Offline stack. Minimal validation: CLINC150 (TF-IDF + multinomial logistic regression trained on the 15,250 training utterances, scored on validation+test, $n = 8{,}600$), CIFAR-100 (the 30,000-image held-out softmax release of Ding et al., 2023), ImageNet validation ($n = 50{,}000$) under torchvision's ResNet-50 (top-1 0.762), GPT-2 on WikiText-2 test ($n = 10{,}000$ positions). Full scale: eight instruction-tuned checkpoints (Qwen2.5-{0.5,1.5,3,7}B, Llama-3.2-{1,3}B, SmolLM2-1.7B, Phi-3.5-mini) scored with fp16/fp32 forward passes over 50,000 positions each of WikiText-103 test at context 512, storing the exact rank of the realized token over the full vocabulary, its probability, the strictly-greater mass, and the sorted top-1024 profile for the first 20,000 positions. Vision: nine timm checkpoints on seeded uniform subsamples of ImageNet-val (six at $n = 20{,}000$, the three largest at $n = 10{,}000$; the timm ResNet-50 of §9.2, top-1 0.80, is a different checkpoint from the torchvision one above), plus {ResNet-50, ConvNeXt-T, ViT-B} $\times$ {gaussian noise, defocus blur, JPEG, brightness} at severity 2, $n = 10{,}000$ each: 270,000 scored images. The validation set is class-sorted; prefix subsampling covers only classes 0–399 and inflates top-1 by $\sim$6 points; we sample uniformly at a fixed seed instead.

The census. Of the 11 identifiers that serve log-probabilities (10 distinct models; the probe file plus the depth sweep of `api_caps.json`, which adds the Qwen routes rejected at depth 20), seven are audited; the other four are `gpt-4.1` and `gpt-4o` (larger siblings of audited routes), `deepseek-v4-flash` (the model `deepseek-chat` itself resolved to during the ablation, so one model under two identifiers) and `qwen-max` (returns a single logprob). Of the other 17 probed identifiers, six answered but refuse informatively (the GPT-5 family rejects `top_logprobs` outright, and `gemini-2.5-flash` returns no logprobs field, though Google documents a cap of 20 on Vertex AI), and eleven could not be reached at probe time (4xx/5xx), so the census is silent about them, Anthropic included, whose absence of logprobs is documented rather than measured here.

Live audit. Seven endpoints queried in August 2026 under the model identifiers of Table 6, every call through one API account and client, so a behaviour we report is that of the route as served there, requesting the per-endpoint maximum `top_logprobs` (5/20/50); temperature 0; one completion token; no other decoding or reasoning parameter set, so `glm-4.7-maas` ran in its default thinking mode (the provider's request-level switch is tested in §B.3); every call checkpointed by index with error rows never marked done (so re-runs never re-pay); per-endpoint concurrency tuned to the failure mode (latency-bound endpoints get more workers, rate-limited ones do not). The text tasks A and C draw seeded 800-character windows at independent random positions of the corpus (so items are not consecutive positions and the Wilson intervals treat them as independent draws), prefix-stable in n, so every lane is the first n items of the largest (verified from the stored truth prefixes: identical on every shared index); the classification lanes launched at $n \in \{300, 800, 1200\}$ on task B, and the n=300 lane on task D, drew different item subsets because `random.sample` is not prefix-stable in n, and this pairing structure, recovered post hoc from the stored truth prefixes, is disclosed in the shipped tables (task E is fully paired across all seven endpoints, which is why the matched-budget reversal is quoted from E). Scores live in label space through the prefix map: the score of the true label is that of the highest-ranked returned token prefixing it ($+\infty$ if none does), where on the classification tasks a token prefixes a label when it does so after case-folding both (the prompt requests lowercase labels) and a whitespace-only token prefixes none, while on the text tasks the match is exact; the stored records keep the exact-case rank, and the analysis recomputes it from the stored window, and a label enters the set when some returned token with score $\leq \hat{q}$ prefixes it, so $Y \in C(X) \iff s^{(k)}(X, Y) \leq \hat{q}$ holds label-wise. *cov@6* is coverage at the deepest depth j whose mean delivered size $\mathbb{E}|S_j|$ is at most six labels (the curves of Figure 3): a budget matched on the population mean, not a per-item truncation, and the finest common budget on task E, since `deepseek-chat`'s first token alone keeps 5.3 labels alive (on task B, where every lane reaches five, the ablation of §B.1 uses cov@5). The understatement factor of §9.3(2) is, per cell, the realized $1 - \mathrm{Acc}_k$ over the mean across calls of $1 - \sum_{r \leq k} \hat{p}_{(r)}$; a cell whose mean implied rate is exactly zero (`deepseek-chat`) has an infinite factor and is excluded from the quoted range, as is `glm-4.7-maas` on classification, whose default reasoning mode scores a preamble token rather than the answer (up to 8,086×, task D; on task E 2,660× in default mode and 16× with thinking off, Table 11). The depth-policy cells of §9.2 are the (profile, α) pairs where both the threshold and the flat policy meet the realized-censoring constraint inside the cached depth horizon (134 of 29 × 5). The value-of-probabilities cells of §6 are the (profile, α) pairs where the rank envelope reaches $1 - \alpha$ inside the horizon and truncated LAC is non-degenerate at some cached k (also 134; LAC is taken at its deepest such k, which is the cached horizon $k = 1{,}000$ in every cell, full access for the vision profiles). Restricted to $k \leq 20$, the 81 cells where truncated LAC is non-degenerate at $k = 20$ and the envelope reaches $1 - \alpha$ within 20 give a median of 1.41×, range 1.02–2.15×, none below one; the 8 sub-unit cells at $k = 1{,}000$ are full-access LAC, which an L1 user could replace by the rank envelope. The depth comparison of §9.1 is also run held out (`depth_heldout.py`: each policy tuned on a random half to realized censoring $\leq \alpha$ and scored on the other half, 50 splits): the threshold policy beats flat in 17 of 19 cells, ties in 1 and loses in 1 (CIFAR-100, α=0.01, −0.3%), saving up to 51%, and nucleus loses to flat in 13 of 17, the in-sample counts. Class-conditional feasibility (`classcond_ci.py`) classifies each class by a one-sided 95% Clopper–Pearson interval on $\mathrm{Recall}_k(y)$; with 50 images per ImageNet class no class is certifiably feasible at α=0.05 for any $k \leq 50$, so ImageNet per-class counts are point estimates only. Lanes that first returned short were re-run to completion against the same checkpoint files, so every cell in Table 6 holds its full launch size; what remains is four individual calls, in three cells, that the provider refuses deterministically (one Wikipedia item rejected with HTTP 400 by both Qwen routes, and two Python items on which Qwen-turbo returned a server error on all eight of eight further attempts at a three-minute timeout), which we drop rather than impute. Ten further calls (`deepseek-chat`: 1 on A, 4 on C; `gpt-4.1-mini`: 1 on C; `gpt-4.1-nano`: 4 on C) returned a well-formed success carrying neither content nor log-probabilities, reproducibly on the same item; these are delivered windows of size zero, in which the truth is censored by construction, and we record them as data — counting them as transport errors would drop the item and flatter Acc_k. `glm-4.7-maas` was launched at $n = 800$ on tasks A–D by design, not by failure, so its task-A lane *is* the common subset: restricting all seven endpoints to the 800 task-A items every one of them completed leaves GLM unchanged and moves the other six upward, by +0.008 to +0.019 (qwen-plus), since the common subset is slightly easier; it changes no ordering. Next-token ground truth is matched by token-prefix against the stored 40-character continuation; the $\geq$ 2-character restatement in §9.3 bounds the single-character credit this rule admits.

Additional tables. The supplementary material ships the complete versions of every table summarized in the text: per-(model, corpus) $k_{\min}$ at five levels; per-corruption vision cells; all 6,380 suite cells with validity, set sizes, size-stratified coverage, depth frontiers, and identification brackets; the five-task seven-endpoint audit grid with Wilson intervals; the interface-delivery audit (per-endpoint $\mathbb{P}(p_1 = 1)$, returned-mass statistics, first-token distributions); the label-set analysis (waste, collision, per-item set sizes, empty-set rates, coverage-vs-size curves); and the strictness restatement of every next-token cell at minimum token lengths 1, 2, 3.

B.1 The output-protocol ablation

Task B re-run on the same 2,000 items with the 151 labels numbered rather than named, everything else fixed (4,000 calls), paired by item (Table 2).

Table 2: Label encoding ablation on task B: named against numbered labels, same 2,000 items, Acc_{20}, mean delivered label-set size, and coverage at a five-label budget.

endpoint	encoding	Acc_{20}	mean $\lvert S_k \rvert$	cov@5 labels
`gpt-4.1-mini`	name	0.828	16.4	**0.785**
`gpt-4.1-mini`	number	**0.872**	63.0	0.694
`gpt-4.1-nano`	name	0.760	14.2	**0.673**
`gpt-4.1-nano`	number	**0.779**	35.9	0.577

The numeric arm wins on raw Acc_{20} (100 items flip to correct against 11 the other way for `gpt-4.1-mini`) and loses at a matched label budget, because a token like `1` keeps every label from 1, 10–19 and 100–151 alive.

B.2 Full result grids

The full grids behind §9: the four minimal label spaces (Table 3) and their multi-split validity check (Table 4), the LLM panel (Table 5), the audit (Table 6), the temperature ablation (Table 7), the live split-conformal run (Table 9) and the live append-label experiment (Table 10). The vision-side $k_{\min}$ values quoted at $\alpha \in \{0.02, 0.01\}$ come from the per-profile suite JSON shipped with the code.

Table 3: The four minimal label spaces of §9.1: top-k accuracy and the population budget $k_{\min}(\alpha)$ at five levels. Bold exceeds the 20-logprob cap; GPT-2 at α=0.01 exceeds the cached k=4,096 horizon. Validity checks use 100 random half splits ($n_{\mathrm{cal}} = n_{\mathrm{test}} = N/2$).

			Acc_k			$k_{\min}(\alpha)$				
dataset	K	N	k=1	5	20	.20	.10	.05	.02	.01
CLINC150	151	8,600	0.843	0.959	0.991	1	2	5	10	19
CIFAR-100	100	30,000	0.597	0.847	0.956	4	9	18	**34**	**49**
ImageNet / ResNet-50	1,000	50,000	0.762	0.929	0.976	2	4	9	**26**	**53**
GPT-2 next token	50,257	10,000	0.343	0.573	0.716	**57**	**282**	**1,021**	**2,873**	>4,096

B.3 September 2026 follow-ups: drift, the thinking switch, and repeated sampling

Drift. Re-issuing the first 120 task-E audit calls on two routes in September 2026 reproduced every covered/censored outcome (Acc_{20} 0.992 and 1.000 on those items, unchanged, 100% agreement item by item). `gpt-4.1-mini` was served under the same dated identifier both times; `deepseek-chat` is now served as `deepseek-flash` (August: `deepseek-v4-flash`), and its top token differs on 21% of the items, 14 of those 25 in capitalization only, which the case-folded rule absorbs. The follow-ups below use the audit's items, prompt and depth; their records ship separately from the 49,696-call audit and do not enter its tables.

The thinking switch. The August audit queried `glm-4.7-maas` in its default thinking mode, in which the first token starts a chain of thought (§9.3(3)). The provider exposes a per-request switch

Table 4: Validity of truncated split conformal on the four minimal label spaces: mean coverage over 100 random half splits ($n_{\text{cal}} = n_{\text{test}} = N/2$ of Table 3), counting censored test points as uncovered. Monte-Carlo standard errors are 0.0002–0.0007; every non-degenerate cell lies within two of them of the band $[1-\alpha,\ 1-\alpha+\frac{1}{n_{\text{cal}}+1}]$. *deg.*: degenerate on every split ($e_k > \alpha$, so $\hat{q} = +\infty$ and the set is all of $\mathcal{Y}$). At these calibration sizes the conditional shortfall of §5 is below 10^{-4} in every non-degenerate cell.

		$\alpha = 0.1$			$\alpha = 0.05$		
dataset	score	k=5	20	100	k=5	20	100
CLINC150	LAC	0.8997	0.8998	0.8998	0.9494	0.9498	0.9498
	APS	0.8993	0.9004	0.9005	0.9494	0.9493	0.9495
CIFAR-100	LAC	*deg.*	0.8997	0.8999	*deg.*	0.9502	0.9500
	APS	*deg.*	0.9000	0.9000	*deg.*	0.9502	0.9499
ImageNet / ResNet-50	LAC	0.8999	0.8997	0.8997	*deg.*	0.9499	0.9499
	APS	0.8999	0.9002	0.9004	*deg.*	0.9501	0.9500
GPT-2 next token	LAC	*deg.*	*deg.*	*deg.*	*deg.*	*deg.*	*deg.*
	APS	*deg.*	*deg.*	*deg.*	*deg.*	*deg.*	*deg.*

Table 5: LLM panel at full vocabulary, 50,000 token positions each (n=2,000 for the finite-sample column). $k_{\min}(\alpha)$ is the population budget, with a block-bootstrap 95% interval over 512-position windows at α=.10; $k_{\min}(\alpha, n, \delta)$ additionally requires degeneracy probability $\le \delta = 0.05$ at calibration size n (§5). Over the sixteen $\alpha \in \{.10, .05\}$ cells the finite-sample budget is 1.23–1.36× the population one and never coincides with it; the ratio grows as α falls, reaching 1.97× at α=.01 (Qwen2.5-3B: 4667/2364).

			Acc_k			$k_{\min}(\alpha)$				$k_{\min}(\alpha, 2000, .05)$		
model	params	V	k=1	20	50	.10	95% CI	.05	.01	.10	.05	.01
`Qwen2.5-0.5B`	0.49B	151,936	0.436	0.806	0.865	102	[94,114]	407	4,215	133	552	6,515
`Llama-3.2-1B`	1.24B	128,256	0.437	0.803	0.865	95	[87,104]	356	3,544	121	465	5,528
`Qwen2.5-1.5B`	1.54B	151,936	0.487	0.850	0.900	50	[45,56]	201	2,522	64	267	4,701
`SmolLM2-1.7B`	1.71B	49,152	0.489	0.854	0.904	46	[42,51]	172	1,821	60	231	3,013
`Qwen2.5-3B`	3.09B	151,936	0.502	0.863	0.909	41	[37,46]	163	2,364	52	220	4,667
`Llama-3.2-3B`	3.21B	128,256	0.457	0.825	0.883	70	[64,76]	257	2,577	89	337	4,123
`Phi-3.5-mini-instruct`	3.82B	32,064	0.550	0.888	0.929	26	[23,29]	93	1,000	32	120	1,605
`Qwen2.5-7B`	7.62B	152,064	0.528	0.882	0.923	30	[26,33]	114	2,180	37	155	4,023

(`enable_thinking`); with it off, the same 1,200 task-E items at the same depth and T=0 return windows that start at the answer (Table 11): $\text{Acc}_1 = 0.897$ and $\text{Acc}_{50} = 0.995$ (default mode: 0.000 and 0.585), $p_{(1)}$ graded (mean 0.928, never exactly 1), and the reported censoring rate understates the realized one 16× rather than 2,660×. Split conformal no longer degenerates at any α tried: truncated LAC returns 3.0, 3.7 and 4.9 labels at $\alpha = 0.2, 0.1, 0.05$ (coverage 0.803, 0.902, 0.951). Once the switch is known this is the deepest L1 window in the panel; without it the same endpoint is L3, and nothing in the response schema tells the two apart.

Repeated sampling (L3) against one truncated call. Su et al. (2024) build conformal sets for API-only models from m sampled answers per input: an answer scores minus its sample frequency, plus weighted terms for the input's normalized entropy and for the answer's semantic similarity to the most frequent answer, and the true answer scores $+\infty$ when it was never sampled. We ran it on the audit's task-E items and prompt: $m = 20$ answers at T=1 per item (one request with n=20 on the OpenAI routes; twenty requests on `deepseek-chat`, which rejects $n > 1$), normalized as in that paper (lowercase, no articles or punctuation) and grouped after removing spaces; similarity from all-MiniLM-L6-v2 embeddings; (λ_1, λ_2) on that paper's grid $[0{:}0.05{:}2]^2$, chosen per (m, α) for the smallest mean set on items 1,000–1,199; every method then runs on items 0–999 over the same 200 random half splits, set sizes in labels (Table 12). Frequency alone ($\lambda = 0$) is within 0.02 labels of the tuned score in every non-degenerate cell. Three readings. (i) The sampled pool is a truncation too: the truth's score is finite only when it was sampled, so Theorem 2 applies with $\text{Acc}^{(m)}$, the share of items whose truth is among the m answers, in place of Acc_k; and it is the harsher truncation: on every route $\text{Acc}^{(20)}$ sits 4–21 points below the window's Acc_{20}, because an answer ranked second at probability 10^{-3} is in the window and almost never in twenty draws. (ii)

Table 6: The complete five-task, seven-endpoint audit. Each cell is n / Acc_k; the min certifiable α is $1 - \mathrm{Acc}_k$. Tasks: A next token (Wikipedia), B CLINC150 (K=151), C next token (Python), D Banking77 (K=77), E DBpedia L3 (K=219). Classification cells (B, D, E) match tokens to labels case-insensitively (§B). All 35 cells are above the 200-call reporting threshold and sum to 49,696, the audit total; the four calls the providers refused deterministically on re-run are itemized in §B.

endpoint	cap	A	B	C	D	E
`gpt-4.1-mini`	20	2000 / 0.949	2000 / 0.828	2000 / 0.974	2000 / 0.959	1200 / 0.988
`gpt-4o-mini`	20	2000 / 0.926	2000 / 0.820	2000 / 0.965	2000 / 0.938	1200 / 0.978
`gpt-4.1-nano`	20	2000 / 0.886	2000 / 0.760	2000 / 0.931	2000 / 0.898	1200 / 0.944
`deepseek-chat`	20	2000 / 0.949	300 / 0.977	300 / 0.920	300 / 0.967	1200 / 0.993
`qwen-plus`	5	1999 / 0.851	1200 / 0.822	1200 / 0.873	1200 / 0.894	1200 / 0.953
`qwen-turbo`	5	1999 / 0.716	1200 / 0.784	1198 / 0.618	1200 / 0.867	1200 / 0.909
`glm-4.7-maas`	50	800 / 0.990	800 / 0.725	800 / 0.993	800 / 0.422	1200 / 0.585

Table 7: The temperature ablation behind §9.3(3): task E, the same 300 items and seed at every temperature (arms paired by item), at each endpoint's served depth k; 3,598 successful calls above the rule. $p_{(1)}$ is the returned top probability, $\mathbb{P}(p_{(1)}{\approx}1)$ the share of calls on which it is within 10^{-6} of 1 (log-probabilities arrive at float32 resolution, so such values sit on about eight levels; 1.0 itself on 24% and 8% of `gpt-4.1-mini` and `qwen-plus` calls at T=0, and on every `deepseek-chat` call at T=0), H the mean entropy of the returned window in nats. `deepseek-chat` returns temperature-scaled probabilities, one-hot at T=0 by construction and spread at T=1; the other three routes do not move beyond Wilson noise. Acc_k is exactly invariant on `deepseek-chat` in August and moves by one item at T=1 in September (temperature scaling preserves the order, so this is serving noise); on the other routes it moves within noise. `glm-4.7-maas`'s arms, 135–154 calls each in August, were completed on the same items in September 2026 (two calls in all are transport failures). [§]Below the rule: the route re-collected in September 2026 (1,200 calls; served as `deepseek-flash` where August's calls were served as `deepseek-v4-flash`), including $T \in \{1.7, 2\}$: the mechanism replicates, and raising T past 1 spreads the window further without reordering it.

endpoint	T	n	k	Acc_k	mean $p_{(1)}$	$\mathbb{P}(p_{(1)}{\approx}1)$	H (nats)
`deepseek-chat`	0.0	300	20	0.997	1.000	**1.000**	0.000
`deepseek-chat`	0.5	300	20	0.997	0.962	**0.153**	0.100
`deepseek-chat`	1.0	300	20	0.997	0.905	**0.000**	0.278
`gpt-4.1-mini`	0.0	300	20	1.000	0.959	0.413	0.105
`gpt-4.1-mini`	0.5	300	20	0.997	0.964	0.400	0.093
`gpt-4.1-mini`	1.0	300	20	0.997	0.965	0.433	0.094
`qwen-plus`	0.0	300	5	0.943	0.960	0.203	0.098
`qwen-plus`	0.5	300	5	0.943	0.958	0.187	0.103
`qwen-plus`	1.0	300	5	0.943	0.956	0.220	0.108
`glm-4.7-maas`	0.0	300	50	0.613	0.727	0.000	0.577
`glm-4.7-maas`	0.5	299	50	0.599	0.844	0.000	0.371
`glm-4.7-maas`	1.0	299	50	0.585	0.720	0.000	0.584
`deepseek-chat`[§]	0.0	300	20	0.997	1.000	**1.000**	0.000
`deepseek-chat`[§]	1.0	300	20	0.993	0.895	**0.000**	0.285
`deepseek-chat`[§]	1.7	300	20	0.997	0.794	**0.000**	0.565
`deepseek-chat`[§]	2.0	300	20	0.997	0.708	**0.000**	0.756

Where it is feasible, sampling resolves what the window cannot: a full answer names one label, while a first token keeps 2.6–5.3 alive, so at α=0.2 sampling returns about one label against the window's 2.3–5.3. (iii) The price is the ceiling, and without n the bill: the window certifies α=0.05 on three of four routes, where sampling degenerates on all four (on 57% of splits for `deepseek-chat`), and `deepseek-chat`'s one-label sets at α=0.1 cost twenty calls and 20× the prompt tokens of one window. On the OpenAI routes n=20 is one request whose prompt is billed once (about 890 tokens against 843).

Table 8: Does the T=1 remedy restore the probabilities? Split conformal on the temperature-ablation calls of Table 7 (task E, 300 items per arm, 200 random half splits, n_{cal}=150), truncated LAC against the truncated rank score, mean label-set size / coverage, sizes in labels as in Table 9. On `deepseek-chat` at T=0 LAC is two-valued on the window: it can do no better than the rank score (α=0.2) and at α=0.1 keeps most of the window; at $T \geq 0.5$ it gives smaller sets than the rank score at target coverage, and T past 1 buys nothing further ([§], the September re-collection). The other routes do not move with T and no cell shown degenerates; at α=0.05 `qwen-plus` ($e_5 = 0.057$ on these items) degenerates on 85% of splits at every T. Every `glm-4.7-maas` cell degenerates.

		$\alpha = 0.2$		$\alpha = 0.1$	
endpoint	T	LAC	rank	LAC	rank
`deepseek-chat`	0.0	5.3 / 0.885	5.3 / 0.885	43.1 / 0.981	6.7 / 0.953
`deepseek-chat`	0.5	4.6 / 0.807	5.1 / 0.889	5.4 / 0.904	6.6 / 0.954
`deepseek-chat`	1.0	4.6 / 0.804	5.3 / 0.885	5.4 / 0.905	6.5 / 0.957
`gpt-4.1-mini`	0.0	2.1 / 0.805	2.5 / 0.865	2.9 / 0.905	3.9 / 0.944
`gpt-4.1-mini`	0.5	2.2 / 0.803	2.4 / 0.846	2.8 / 0.904	4.0 / 0.945
`gpt-4.1-mini`	1.0	2.2 / 0.802	2.4 / 0.878	2.8 / 0.905	4.0 / 0.934
`qwen-plus`	0.0	4.0 / 0.811	4.3 / 0.856	5.3 / 0.904	6.8 / 0.916
`qwen-plus`	0.5	3.8 / 0.809	4.2 / 0.879	5.2 / 0.905	6.4 / 0.918
`qwen-plus`	1.0	3.9 / 0.806	4.3 / 0.862	5.2 / 0.903	6.6 / 0.919
`deepseek-chat`[§]	0.0	5.1 / 0.911	5.1 / 0.911	18.2 / 0.935	5.4 / 0.919
`deepseek-chat`[§]	1.0	4.5 / 0.805	5.0 / 0.908	5.1 / 0.903	5.4 / 0.918
`deepseek-chat`[§]	1.7	4.5 / 0.804	5.1 / 0.905	5.3 / 0.903	5.6 / 0.923
`deepseek-chat`[§]	2.0	4.7 / 0.805	5.3 / 0.907	5.4 / 0.902	5.6 / 0.917

B.4 SEPTEMBER 2026 GPU FOLLOW-UPS: REPRODUCTION, BASE CHECKPOINTS, A TRAINING-CUTOFF CONTROL, AND THE LABEL SPACE AT FIXED WEIGHTS

Reproduction. The eight-checkpoint panel and the 21 vision cells were re-scored on a different GPU and software build (one 24 GB GPU; torch 2.14, transformers 5.17, timm 1.0.30), the two checkpoints scored in fp32 on CPU in August again in fp32. Every vision cell reproduces its shipped $k_{\min}(\alpha)$ at all five levels, except two cells that move by one at α=0.01 (116 to 115 and 181 to 180), and its top-1 accuracy to within two images. Every LLM reproduces $k_{\min}(\alpha)$ at $\alpha \in \{0.2, 0.1\}$, five of eight at all five levels, and the other three within 1 at α=0.05 and 0.3% at α=0.01 (fp16 kernels differ across builds; Acc_1 moves by at most 4×10^{-5}). The block-bootstrap intervals of Table 5 are recomputed by the shipped `kmin_ci.py` (1,000 resamples of 512-position windows) and move by at most 2 at either end. The conformal suite recomputed from the regenerated caches returns the same degeneracy status in all 6,380 cells, non-degenerate coverages within 0.0011 of the shipped ones, and the same summary statistics of §9.2 (mean deviation +0.0013, sd 0.0064, worst cell −0.017, unconditional +0.0220; depth policy 118/9/7 of 134; value of probabilities 1.50×, 0.42–2.35×), except that 4.3% rather than 4.4% of cells fall more than a point below target.

Instruction tuning and the training cutoff (Table 13). The panel scores instruction-tuned checkpoints on WikiText-103, text every model has plausibly seen in training. We add the base checkpoint of every family that publishes one (seven) and two corpora from one outlet on either side of every model's cutoff: BBC News articles from January–April 2023 and from March–June 2025 (`RealTimeData/bbc_news_alltime`; lines repeated five or more times, i.e. video notices and newsletter plugs, dropped; articles in a seeded random order; 50,000 positions each, as in the panel). Base checkpoints need fewer logprobs than their instruction-tuned twins on WikiText ($k_{\min}(0.1)$ smaller by 1.04–1.84×; Llama-3.2-3B 38 against 70), so the panel's budgets are those of the checkpoints an API serves, not of the best next-token predictors. Fresh text needs more: on 2025 news $k_{\min}(0.1)$ is 1.23–1.50× its 2023 value for every one of the fifteen checkpoints, with disjoint intervals in every case, so a frontier measured on pre-cutoff text is optimistic for the text a deployed model will see. Neither rescues the cap: the smallest $k_{\min}(0.1)$ in the 45 (checkpoint, corpus) cells is 25 (Phi-3.5-mini, 2023 news), and on 2025 news every checkpoint but Phi-3.5-mini (32) needs 38 or more.

Table 9: Split conformal run on the live task-E calls (DBpedia, K=219, 1,200 items on every one of the seven endpoints; 200 random splits into $n_{\text{cal}} = n_{\text{test}} = 600$). Every cell is the mean over all 200 splits, a degenerate split counted at coverage 1 and size 219. *: partially degenerate cells, at the transition of Theorem 2: `gpt-4.1-nano` at α=0.05 (67 censored of 1,200, $e_{20} = 0.056$) degenerates on 82.5% of splits, `qwen-plus` at α=0.05 (56 censored, $e_5 = 0.047$, inside the feasible region) on 37.0%, `qwen-turbo` at α=0.1 (109 censored, $e_5 = 0.091$) on 11.0%; the hypergeometric law of a random split predicts 84.3%, 34.1% and 15.8%, where the binomial law of an i.i.d. draw would say 75.7%, 37.6% and 23.6%. On their non-degenerate splits these cells give 19.7, 7.3 and 7.1 labels (LAC) at 0.931, 0.945 and 0.896, below target as marginal validity permits; the conditional law of §5 predicts 0.930, 0.943 and 0.895 (Appendix A). *LAC* needs the returned probabilities (L1); *rank* needs only their order (L2). **Sizes are labels, not token slots**: a label is in the set when some returned token with score $\leq \hat{q}$ prefixes it (case-insensitively, §B), so these are comparable across endpoints in the sense §9.3 requires. Degeneracy is flagged by the $\hat{q} = +\infty$ criterion of Theorem 2, at which the truncated set is all of $\mathcal{Y}$ and the size is K=219 — *not* k. On the endpoints that deliver L1, LAC never returns the larger set. `deepseek-chat` is the case the diagnostics predict: its probabilities are constant, so LAC is two-valued on the window and its set is either the first token's labels (5.3) or the whole window (49.3); at α=0.1 the calibration quantile falls on one side or the other from split to split (26.7 on average), and at α=0.05 LAC keeps the window (49.4) while the rank score, which needs strictly less access, delivers 7.9 labels at 95.9% coverage — a 6.3× reduction from discarding the probabilities.

		$\alpha = 0.2$		$\alpha = 0.1$		$\alpha = 0.05$	
endpoint	score	cov	size	cov	size	cov	size
`gpt-4.1-mini`	lac	0.802	2.3	0.901	3.0	0.950	3.7
	rank	0.870	2.7	0.934	4.2	0.960	5.6
`gpt-4o-mini`	lac	0.801	2.6	0.901	4.0	0.950	8.2
	rank	0.822	2.7	0.915	5.0	0.952	8.9
`gpt-4.1-nano`	lac	0.802	4.8	0.902	11.2	0.988	184.1*
	rank	0.821	6.2	0.903	13.1	0.988	184.3*
`deepseek-chat`	lac	0.901	5.3	0.942	26.7	0.993	49.4
	rank	0.901	5.3	0.924	6.1	0.959	7.9
`qwen-plus`	lac	0.801	3.6	0.901	4.8	0.965	85.6*
	rank	0.877	4.1	0.915	6.2	0.967	86.2*
`qwen-turbo`	lac	0.800	3.4	0.908	30.4*	*1.000*	*219.0*
	rank	0.837	5.1	0.913	31.7*	*1.000*	*219.0*
`glm-4.7-maas`	lac	*1.000*	*219.0*	*1.000*	*219.0*	*1.000*	*219.0*
	rank	*1.000*	*219.0*	*1.000*	*219.0*	*1.000*	*219.0*

Italic = degenerate on every split ($\hat{q} = +\infty$; set = $\mathcal{Y}$, size K=219); * = on some splits.

The label space at fixed weights. Across the panel, vocabulary is collinear with family (§9.2). Merging, for the same weights, text and positions, every set of tokens whose decoded strings agree after stripping whitespace and case-folding into one class (probabilities summed) shrinks the label space by 26–32% (Qwen2.5: 151,936 to 113,135 classes) and lowers $k_{\min}(0.1)$ by 7–14% on WikiText and by 0–5% on news (Table 13, last columns): a real but second-order axis, in line with the within-noise matched pairs of §9.2.

Vision intervals (Table 14). The vision $k_{\min}$ values of §9.2 carried no uncertainty; images are an i.i.d. sample of ImageNet-val, so we bootstrap them (1,000 resamples, `kmin_ci.py`). Both vision statements survive their intervals: the two strongest checkpoints need at most 15 logprobs at α=0.01 (upper ends 15 and 14, inside the 20-cap), and defocus blur's $k_{\min}(0.05)$ for ResNet-50 is at least 48, beyond it.

Table 10: One appended label, live (task B, CLINC150, K=151, at each endpoint's served depth): the label is the most frequent censored truth on a calibration half and is scored on the disjoint test half, 200 seeded half splits; *recovered* is the test-half gain over the test-half deficit $1 - \mathrm{Acc}_k$, mean $\pm$ sd over splits. The mined label is the out-of-scope intent `oos` on every split of every lane, so the recovery is that intent's share of the censored items; `glm-4.7-maas`'s window holds preamble rather than answers. Lanes differ in n and, for $n < 2{,}000$ on this task, in item subset (§B).

endpoint	n	Acc_k	deficit	recovered
`deepseek-chat`	300	0.977	0.023	100% $\pm$ 0.0
`gpt-4.1-mini`	2000	0.828	0.173	93% $\pm$ 1.3
`gpt-4o-mini`	2000	0.820	0.181	89% $\pm$ 1.6
`qwen-plus`	1200	0.822	0.178	86% $\pm$ 2.3
`qwen-turbo`	1200	0.784	0.216	78% $\pm$ 2.5
`gpt-4.1-nano`	2000	0.760	0.240	74% $\pm$ 2.2
`glm-4.7-maas`	800	0.725	0.275	16% $\pm$ 2.7

Table 11: `glm-4.7-maas` on task E with and without its per-request thinking switch: the same 1,200 items, depth 50, T=0 (August: default mode; September 2026: `enable_thinking` off). $|S_{50}|$ and cov@6 as in Table 1; LAC is truncated split conformal as in Table 9, size / coverage, *italic* degenerate.

mode	Acc_1	Acc_{20}	Acc_{50}	$\lvert S_{50}\rvert$ / cov@6	understatement	LAC, α=0.1	LAC, α=0.05
default (thinking)	0.000	0.244	0.585	28.1 / 0.108	2,660×	*219.0 / 1.000*	*219.0 / 1.000*
thinking off	0.897	0.990	0.995	52.0 / 0.948	16×	3.7 / 0.902	4.9 / 0.951

Table 12: One truncated call against m sampled answers (Su et al., 2024) on the same 1,000 task-E items and 200 half splits (n_{cal}=500): mean label-set size / coverage. *window*: the audit's T=0 call at the served depth (k=20), truncated LAC or rank score, labels through the case-folded prefix map; *sampling*: the tuned score of Su et al. (2024) on m answers at T=1 (the first m of the twenty collected). Calls and tokens are per item (prompt plus completion; for m=5 prorated from the m=20 run). The ceiling is Acc_{20} for the window (every label a returned token keeps alive, paid for in set size) and $\mathrm{Acc}^{(m)}$ for sampling (an exact normalized match), so the two ceilings differ partly by rule; the matched-α sets are the like-for-like comparison. *Italic*: degenerate on every split (size 219); *: on some splits.

endpoint	method	calls	tokens	ceiling	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$
`gpt-4.1-mini`	window, LAC	1	843	0.988	2.3 / 0.801	2.9 / 0.899	3.6 / 0.950
	window, rank	1	843	0.988	2.7 / 0.871	4.3 / 0.935	5.6 / 0.961
	sampling, m=5	1	854	0.885	0.9 / 0.815	214.6 / 0.997*	*219.0 / 1.000*
	sampling, m=20	1	890	0.910	0.9 / 0.812	34.9 / 0.911*	*219.0 / 1.000*
`gpt-4o-mini`	window, LAC	1	843	0.979	2.6 / 0.800	3.9 / 0.901	8.0 / 0.951
	window, rank	1	843	0.979	2.8 / 0.822	4.9 / 0.915	8.3 / 0.952
	sampling, m=5	1	854	0.831	2.1 / 0.803*	*219.0 / 1.000*	*219.0 / 1.000*
	sampling, m=20	1	891	0.856	1.0 / 0.802	*219.0 / 1.000*	*219.0 / 1.000*
`gpt-4.1-nano`	window, LAC	1	843	0.946	4.8 / 0.798	11.2 / 0.900	176.1 / 0.985*
	window, rank	1	843	0.946	6.3 / 0.817	13.1 / 0.902	176.4 / 0.985*
	sampling, m=5	1	854	0.683	*219.0 / 1.000*	*219.0 / 1.000*	*219.0 / 1.000*
	sampling, m=20	1	889	0.736	*219.0 / 1.000*	*219.0 / 1.000*	*219.0 / 1.000*
`deepseek-chat`	window, LAC	1	868	0.992	5.3 / 0.904	21.0 / 0.932	49.4 / 0.992
	window, rank	1	868	0.992	5.3 / 0.904	6.0 / 0.920	7.5 / 0.959
	sampling, m=5	5	4344	0.935	0.9 / 0.857	1.0 / 0.905	214.6 / 0.998*
	sampling, m=20	20	17375	0.949	0.9 / 0.818	1.0 / 0.900	126.4 / 0.974*

Table 13: $k_{\min}(0.1)$ with a block-bootstrap 95% interval (1,000 resamples of 512-position windows), 50,000 positions per cell, for instruction-tuned and base checkpoints on WikiText-103 and on BBC News before (Jan–Apr 2023) and after (Mar–Jun 2025) every model's training cutoff; *ratio* is 2025 over 2023. Last two columns: the same WikiText positions in the merged label space (tokens equal up to case and surrounding whitespace summed into one class), its size and $k_{\min}(0.1)$. Phi-3.5 has no public base checkpoint.

			$k_{\min}(0.1)$ [95% CI]				merged, WikiText	
model	variant	V	WikiText	BBC 2023	BBC 2025	ratio	classes	$k_{\min}(0.1)$
`Qwen2.5-0.5B`	instruct	151,936	102 [94,114]	75 [69,81]	103 [92,114]	1.37	113,135	88
	base	151,936	92 [83,101]	70 [64,76]	95 [85,106]	1.36	113,135	81
`Llama-3.2-1B`	instruct	128,256	95 [87,104]	62 [57,66]	76 [69,84]	1.23	87,410	83
	base	128,256	59 [55,66]	38 [35,40]	49 [45,53]	1.29	87,410	53
`Qwen2.5-1.5B`	instruct	151,936	50 [45,56]	44 [40,47]	59 [53,65]	1.34	113,135	45
	base	151,936	48 [44,53]	41 [38,45]	58 [51,63]	1.41	113,135	44
`SmolLM2-1.7B`	instruct	49,152	46 [42,51]	34 [32,36]	51 [45,56]	1.50	33,181	42
	base	49,152	39 [35,44]	31 [28,33]	43 [40,47]	1.39	33,181	35
`Qwen2.5-3B`	instruct	151,936	41 [37,46]	36 [34,39]	48 [44,53]	1.33	113,135	37
	base	151,936	38 [34,42]	34 [32,37]	46 [41,51]	1.35	113,135	35
`Llama-3.2-3B`	instruct	128,256	70 [64,76]	41 [39,44]	55 [50,61]	1.34	87,410	62
	base	128,256	38 [36,42]	28 [26,30]	38 [34,41]	1.36	87,410	35
`Phi-3.5-mini`	instruct	32,064	26 [23,29]	25 [23,27]	32 [29,35]	1.28	22,280	24
`Qwen2.5-7B`	instruct	152,064	30 [26,33]	30 [28,33]	41 [37,44]	1.37	113,263	27
	base	152,064	27 [25,30]	29 [27,32]	38 [35,41]	1.31	113,263	25

Table 14: The 21 vision cells of §9.2: $k_{\min}(\alpha)$ with a 95% bootstrap interval over images (1,000 resamples) at $\alpha \in \{0.05, 0.01\}$; bold exceeds the 20-logprob cap. Corruptions at severity 2. †: the September re-run gives a point value one smaller (115, 180), on which its interval is computed.

model	input	n	top-1	$k_{\min}(0.1)$	$k_{\min}(0.05)$ [95% CI]	$k_{\min}(0.02)$	$k_{\min}(0.01)$ [95% CI]
`mobilenetv3-small`	clean	20,000	0.680	7	19 [18,21]	**63**	**134** [117,152]
`resnet18`	clean	20,000	0.718	6	15 [14,16]	**46**	**92** [81,105]
`efficientnet_b0`	clean	20,000	0.778	3	7 [7,8]	**22**	**56** [47,61]
`resnet50`	clean	20,000	0.802	3	6 [6,7]	**21**	**50** [43,60]
`resnet50`	noise	10,000	0.752	4	12 [11,14]	**42**	**104** [83,129]
`resnet50`	defocus	10,000	0.595	17	**56** [48,64]	**156**	**265** [238,295]
`resnet50`	JPEG	10,000	0.672	9	**26** [23,30]	**93**	**181** [152,202]†
`resnet50`	brightness	10,000	0.778	3	9 [8,11]	**31**	**71** [59,89]
`convnext_tiny`	clean	20,000	0.821	2	4 [4,5]	12	**26** [22,34]
`convnext_tiny`	noise	10,000	0.790	3	6 [6,7]	**22**	**53** [42,72]
`convnext_tiny`	defocus	10,000	0.661	9	**25** [23,29]	**102**	**222** [186,270]
`convnext_tiny`	JPEG	10,000	0.721	5	14 [13,16]	**49**	**116** [91,186]†
`convnext_tiny`	brightness	10,000	0.799	3	6 [5,6]	17	**39** [31,49]
`deit3_small`	clean	20,000	0.815	2	5 [5,5]	18	**55** [44,68]
`swin_tiny`	clean	10,000	0.822	2	5 [4,5]	15	**32** [26,42]
`vit_base`	clean	10,000	0.859	2	3 [3,3]	6	11 [10,15]
`vit_base`	noise	10,000	0.837	2	4 [3,4]	9	20 [15,24]
`vit_base`	defocus	10,000	0.789	3	6 [6,7]	20	**47** [41,64]
`vit_base`	JPEG	10,000	0.802	3	5 [5,6]	16	**37** [30,47]
`vit_base`	brightness	10,000	0.839	2	3 [3,4]	8	17 [14,20]
`convnext_base`	clean	10,000	0.864	2	3 [3,3]	6	11 [10,14]