Conformal Selective Acting: Anytime-Valid Risk Control for RLVR-Trained LLMs

Hamed Khosravi **Xiaoming Huo** Georgia Institute of Technology

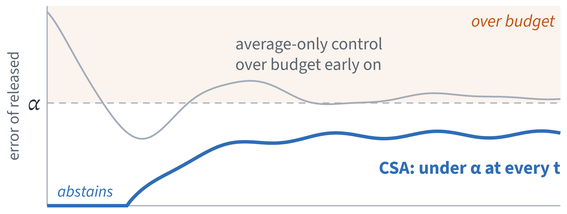
A learning LLM releases answers under **certified thresholds**, with high-probability, **anytime-valid** control of selective risk on each deployment stream.

NeurIPS 2026 · Main Track

1 The deployment problem

A hospital runs its **own specialist LLM** — and keeps improving it on the job.

- **Keeps learning:** policy updated by RLVR
- **Verifier:** each answer graded right or wrong
- **Release or abstain:** uncertified answers are withheld
- **Promise:** errors among released answers ≤ α, at **every** moment, on **this one** stream



Schematic. A guarantee that only holds on average, or in the long run, can be over budget for long stretches of the one stream that is actually deployed.

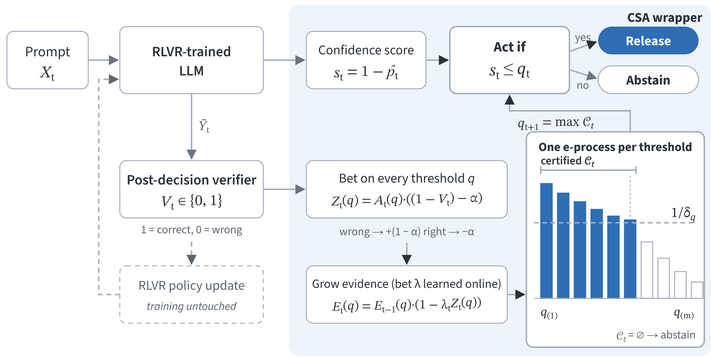
2 Why existing guarantees fall short

Guarantee	Selective risk	Every round	Updated policy
Fixed-horizon RCPS, LTT, CRC, ConfFact	✓	✗	✗
Long-run average ACI, SAOCP	✗	✗	✓
Anytime, marginal risk A-RCPS	✗	✓	✓
CSA (ours) anytime-pathwise, selective	✓	✓	✓

Selective risk: error rate counted over **released** answers only.

3 Conformal Selective Acting

A wrapper around the RLVR loop. Every candidate threshold collects its own **betting evidence** that it is safe; CSA acts under the **loosest certified** one.



4 What CSA guarantees

Target. Error rate among released answers up to round T:

$$R_T^{\text{act}} = \frac{\Sigma_{t \le T} A_t (1 - V_t)}{\Sigma_{t \le T} A_t = N_T}$$

$A_t = 1$: answer released · $V_t = 1$: verifier marks it correct · N_T: releases so far

Theorem 5.1 · Anytime selective-risk bound

with prob. ≥ 1 − 2δ (δ: failure level you choose), simultaneously for all T

$$R_T^{\text{act}} \le \alpha + \bar{\nu}_T + c_1 \sqrt{\frac{L_T}{N_T \vee 1}} + c_2 \frac{L_T}{N_T \vee 1}$$

α — the budget you choose

$\bar{\nu}_T$ — drift slack: how far risk can move between policy updates

c_1, c_2 — absolute constants; $L_T = \log(1/\delta) + \log\log N_T$; these terms shrink as releases N_T grow

Assumptions. predictability · monotone risk in q · frontier stability · drift-budget dominance. No exchangeability needed.

Theorems 5.3–5.4 · Optimal speed

$$\mathbb{E}[\tau_{j,q}^{\text{cert}} - \tau_j] \le \frac{4\,(\log(1/\delta_{j,q}) + 1)}{(\bar{\eta}_{j,q} - \nu_j)^2}$$

τ_j: start of epoch j (rounds between policy updates) · $\tau^{\text{cert}}_{j,q}$: round threshold q gets certified · $\bar{\eta}_{j,q}$: safety margin of q (α minus its risk) · ν_j: drift within the epoch · $\delta_{j,q}$: share of δ given to q

Certification delay with an **oracle fixed bet**; matches the lower bound $\Omega(\log(1/\delta) / \bar{\eta}^2)$ — **rate-optimal**.

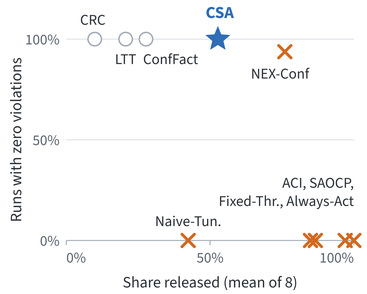
Theorem 5.5 · Horizon-free missed-release gap

Cumulative missed-release gap to the safe-frontier oracle is bounded **independently of** T, under the theorem's conditions.

5 Evidence

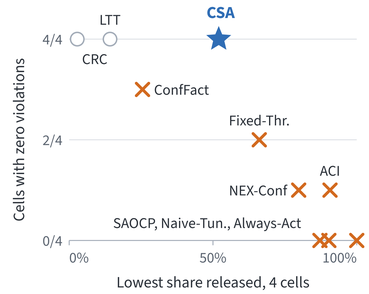
★ **CSA (ours)** ○ never violated ✕ violated in at least one run

8 specialist benchmarks



Pivotal α per benchmark, 10 runs each. Medicine, finance, law, science, math.

Live RLVR, online LoRA



Expert-Iteration RLVR, policy updated during deployment. 4 cells × 20 reps at tightest α*. Four newer live-RLVR cells shown; fifth (MATH) cell reported in the paper.

Top-right is the goal — never over budget, still answering. Only CSA gets there in both settings; safe baselines buy safety by refusing.

Safety vs. abstention across risk budgets

Share of answers CSA releases; 48 cells, **0 pathwise violations in every cell**

	α = 0.05	α = 0.10	α = 0.15	α = 0.20	α = 0.25	α = 0.30	base error
MedQA	*abstain*	*abstain*	29%	39%	64%	82%	31.5%
PubMedQA	*abstain*	*abstain*	*abstain*	63%	92%	96%	23.9%
TAT-QA	*abstain*	*abstain*	46%	50%	51%	93%	25.7%
MedNLI	*abstain*	19%	25%	29%	97%	99%	21.0%
GSM8K	72%	94%	99%	100%	100%	100%	5.0%
HEAD-QA	*abstain*	*abstain*	*abstain*	62%	82%	95%	26.0%
ARC-C	*abstain*	56%	94%	96%	96%	97%	10.0%
CaseHOLD	*abstain*	*abstain*	*abstain*	23%	50%	74%	34.0%

When no threshold can be certified at a tight α, CSA **abstains** rather than gamble; as α loosens, it releases more.

0 / 740
CSA runs with a pathwise violation, across benchmarks, shift and live RLVR

52.7%
mean share released at the pivotal α; zero-violation baselines reach ≤ 27.7%

4 / 4
live RLVR cells safe while releasing ≥ 52% — no baseline does both

Paper Code LinkedIn