*Robustness → Antifragility**: Bridging Design-Time Safety & Test-Time Adaptation

Power Systems, Cybersecurity, and beyond

Ming Jin

VIRGINIA TECH

Antifragility: “Systems that benefit from shocks and volatility, emerging stronger over time.” - N. Taleb

Partners: NREL, GE Global Research, Southern Company

Sponsors: NSF, Commonwealth Cyber Initiative Coastal Virginia, Deloitte.

From Known Risks to Unknown Threats

Antifragility = Bridging Offline Safety & Online Adaptation

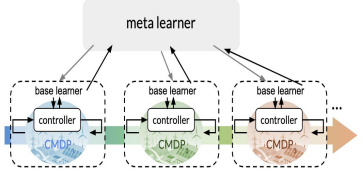
Time

Run-Time (Online)

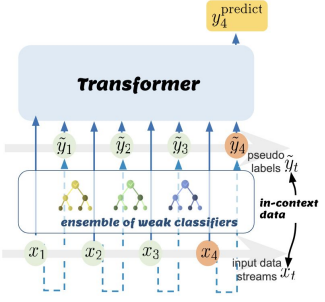
Design-Time (Offline)

Meta Safe RL

multi-task → quick adaptation



In-Context Detection **w/o** Feedback



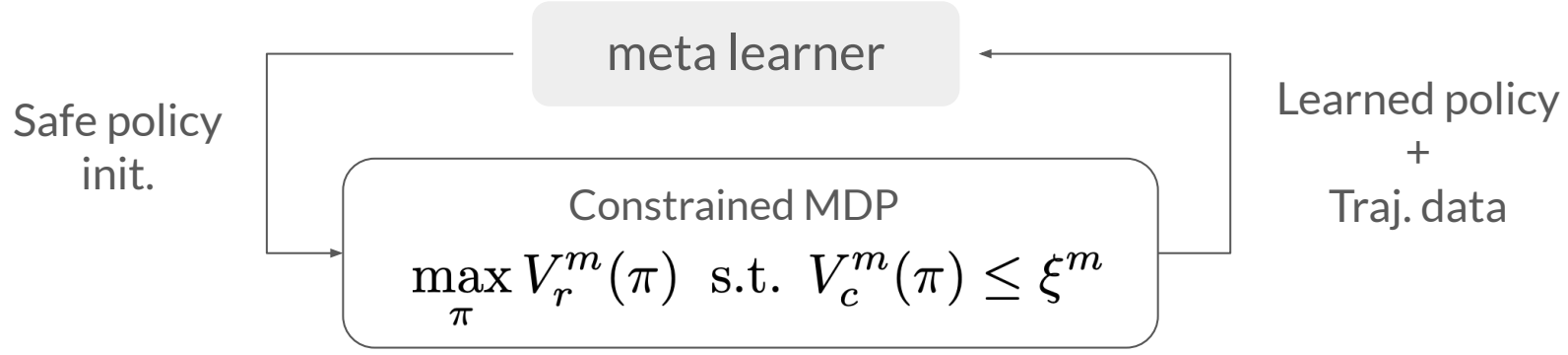
Robust control/
Classical safe RL /
Pre-deployment verification

Known/Anticipated Risks
(*e.g.* traditional safety guarantees)

Unknown/Unexpected Risks
(*e.g.* OOD failures, zero-day attacks)

Certainty

Meta-Safe RL: Learning to do Safe RL *Fast*



Goal: adapt the **policy initialization** to minimize:

Task-averaged:

optimality gap $\overline{R}_r = \frac{1}{M}\sum_{m=1}^{M}[V_r^m(\pi^{*m}) - V_r^m(\hat{\pi}^m)]$

constraint violation $\overline{R}_c = \frac{1}{M}\sum_{m=1}^{M}[V_c^m(\hat{\pi}^m) - \xi^m]$

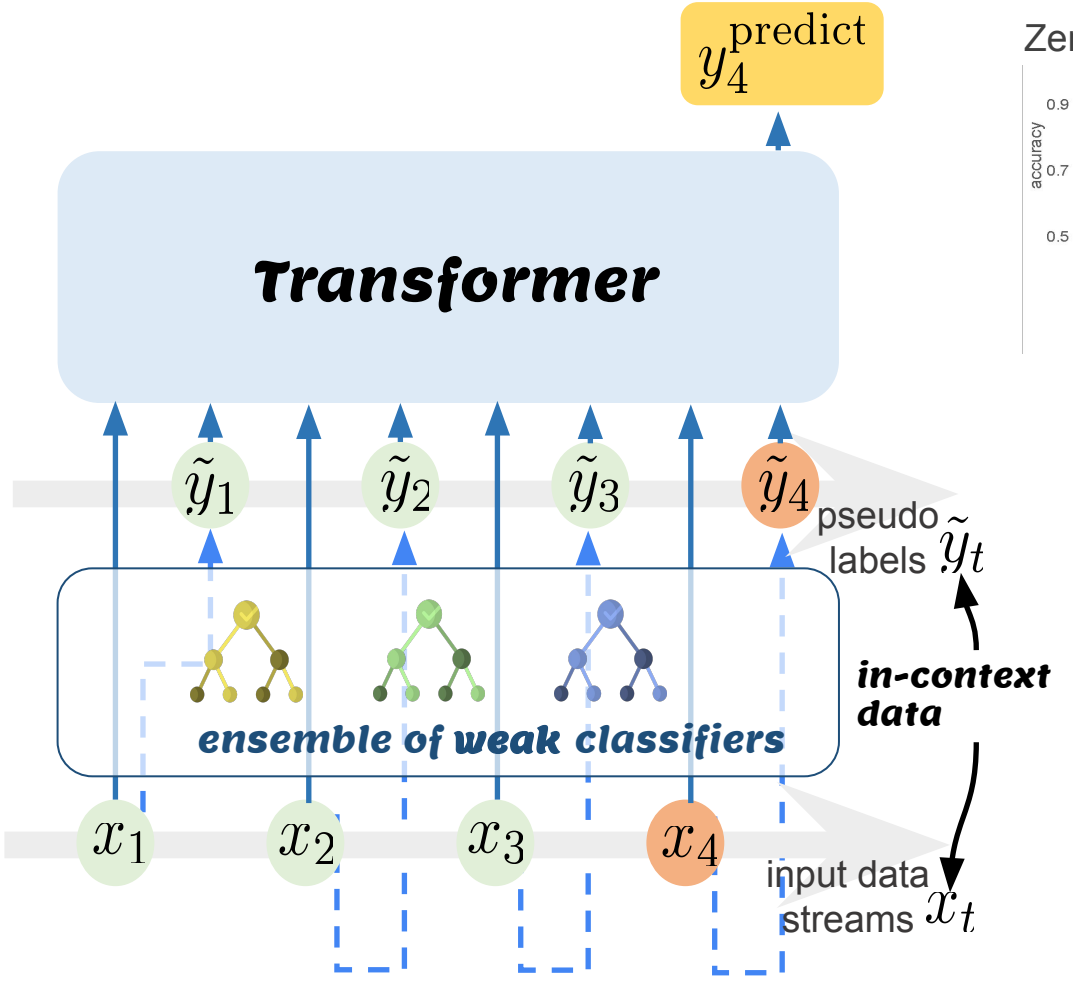
Adaptive safety bound

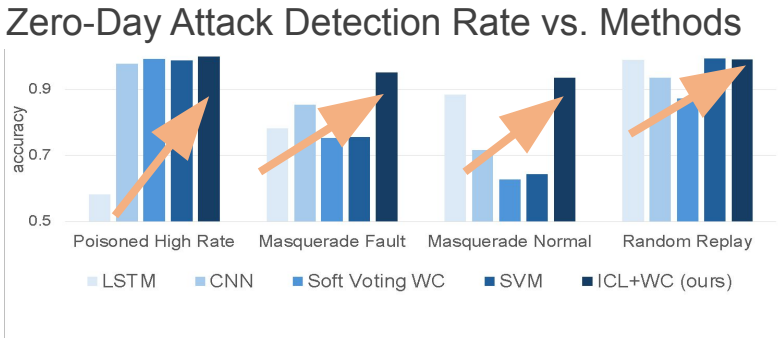
$$\mathcal{O}\left(\frac{\text{Env. variations}}{\sqrt{\#\text{ steps}\#\text{ past CMDPs}}}\right)$$

A CMDP-within-online framework for Meta-Safe Reinforcement Learning. Khattar, V.; Ding, Y.; Sel, B.; Lavaei, J.; and Jin, M. ICLR 2023. (spotlight presentation)

Applications: critical load restoration (w/ NREL), automated pen-testing (Deloitte's RASOR platform)

Zero-Day ICS Attacks: In-Context Detection *W/O* Feedback





- **Challenge:** No labeled data or real-time feedback for novel attacks.
- **Method:** Pretrained transformer + minimal heuristics (weak classifiers) → in-context labels, no fine-tuning.
- **Result:** ~85% detection on ICS data

CPS demo (2024 Oct.)

Field test (2025 Q4)

A Roadmap to Safe & Antifragile AI

Towards Antifragility:
use near-/imagined-failures as feedback, minimize labeled data reliance , in-context safe RL, …

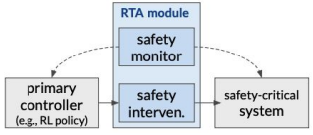
Time

Run-Time (Online)

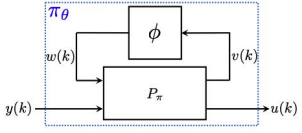
Design-Time (Offline)

Multi-Objective / Safe RL

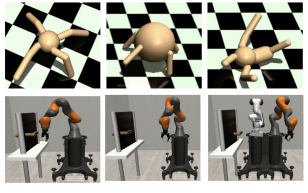
Run-Time Assurance



Robust Control



Robust Testing



Safe Policy → Safe Adaptation Process

Known/Anticipated Risks
(*e.g.* traditional safety guarantees)

Unknown/Unexpected Risks
(*e.g.* OOD failures, zero-day attacks)

Certainty