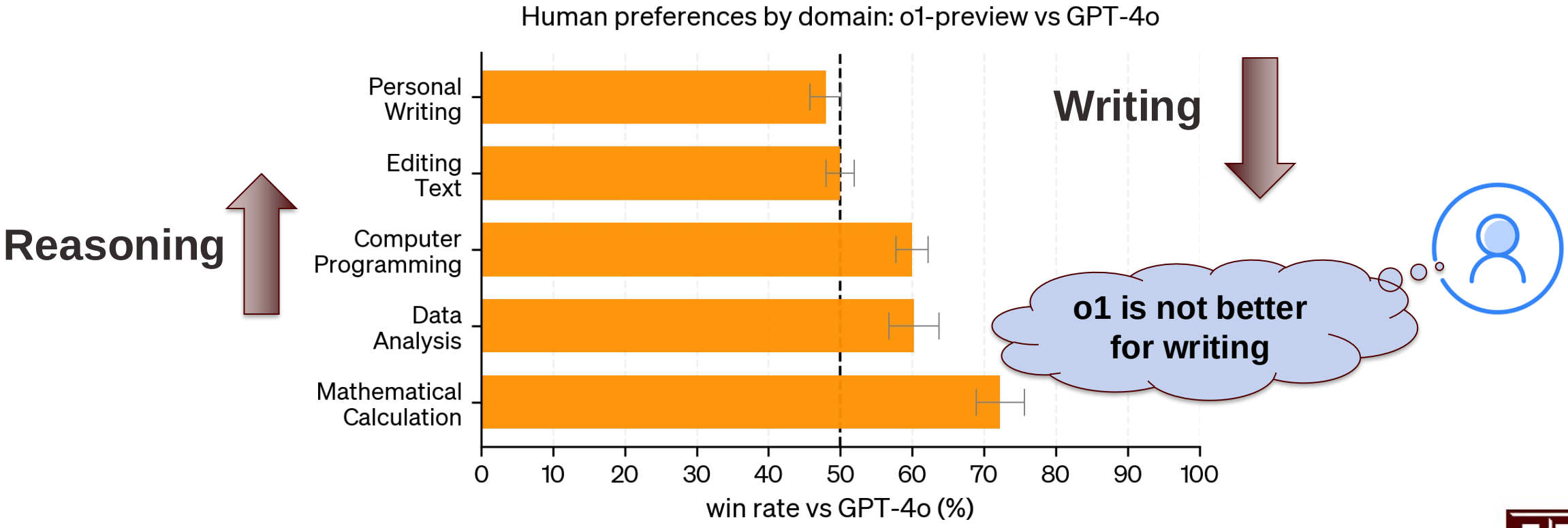
Model Developmental Safety

Tianbao Yang
CSE@TAMU

Continual Model Development

GPT-3.5 → GPT-4 → GPT-4o → o1 →...



Catastrophic Forgetting

Cost-sensitive Applications

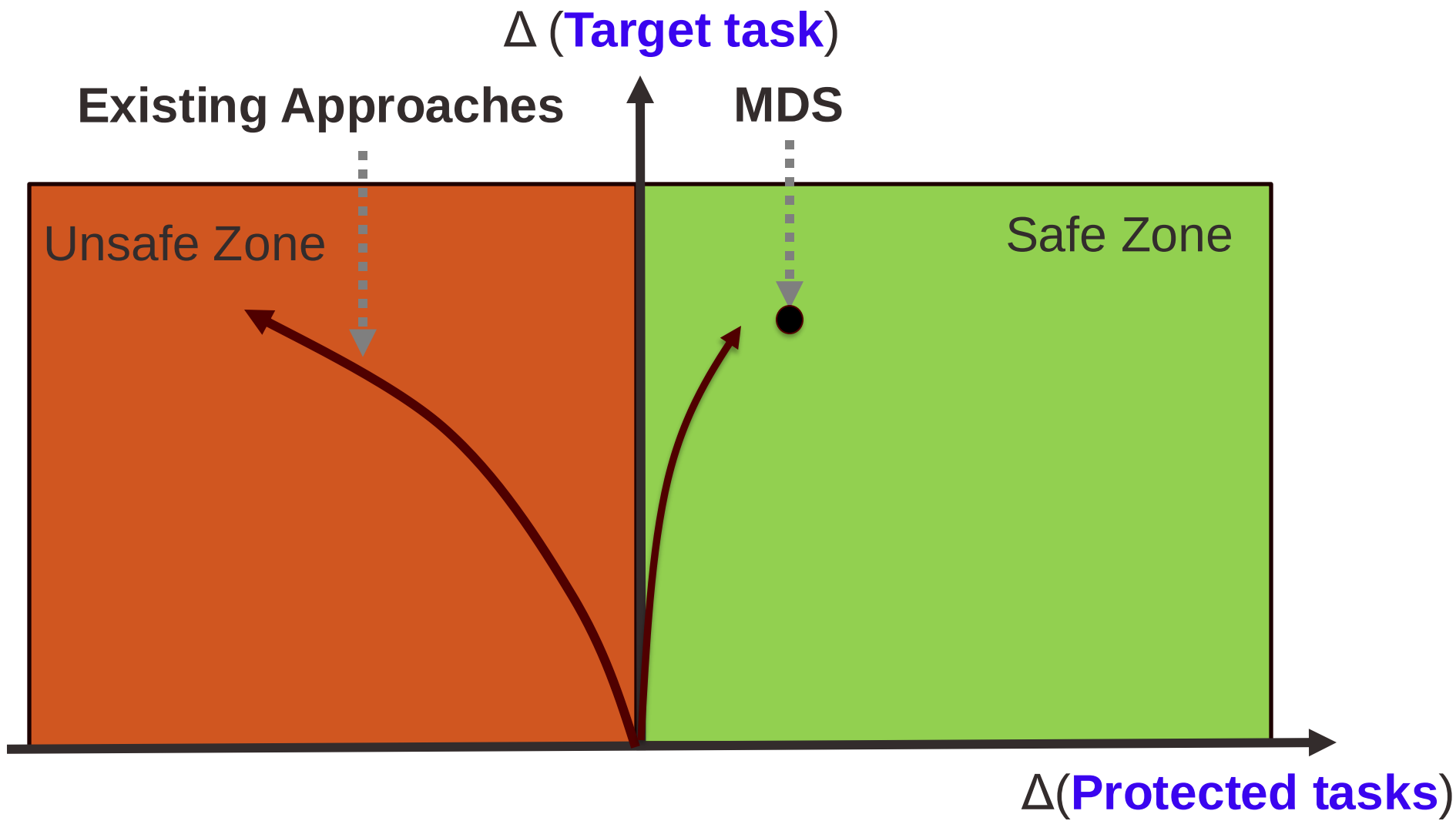
Improving ***Foggy*** may degrade ***Clear***

Re-validation requires extensive **costs**

Improving ***common diseases*** may degrade ***rare ones***

Cause **loss of life**

Model Developmental Safety (MDS)



Mathematical Framework

Non-convex Constrained Optimization

$$\mathbf{w}_{\text{new}} = \arg\min_{\mathbf{w}} F(\mathbf{w})$$

Target Task Objective

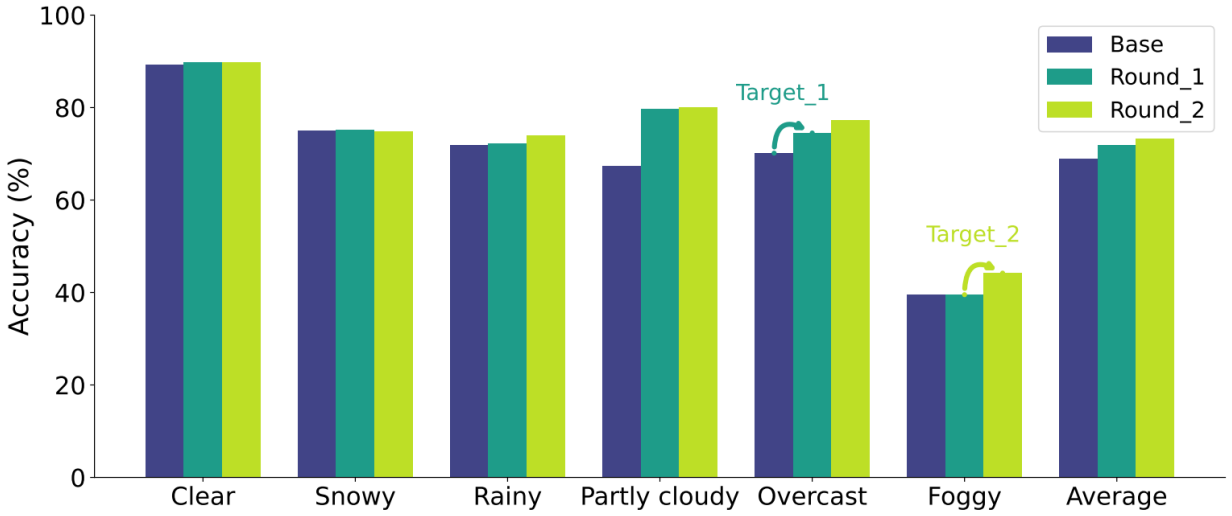
$$s.t. \quad L_k(\mathbf{w}) \leq L_k(\mathbf{w}_{\text{old}}), \quad k = 1, \ldots, m$$

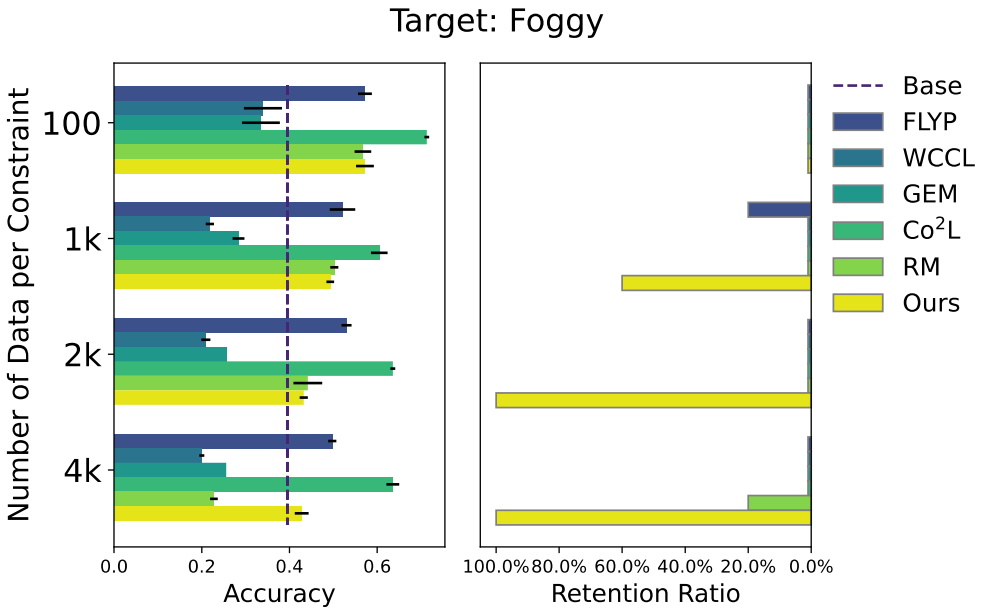
Protected task Loss

Squared-Hinge Penalty Method (Provable Convergence)

$$\min_{\mathbf{w}} F(\mathbf{w}) + \frac{\beta}{m} \sum_{k=1}^{m} [L_k(\mathbf{w}) - L_k(\mathbf{w}_{\text{old}})]_+^2$$

Autonomous Driving





Acknowledgements: NAIRR Award 240040 and NCSA Delta

Please come to our poster!