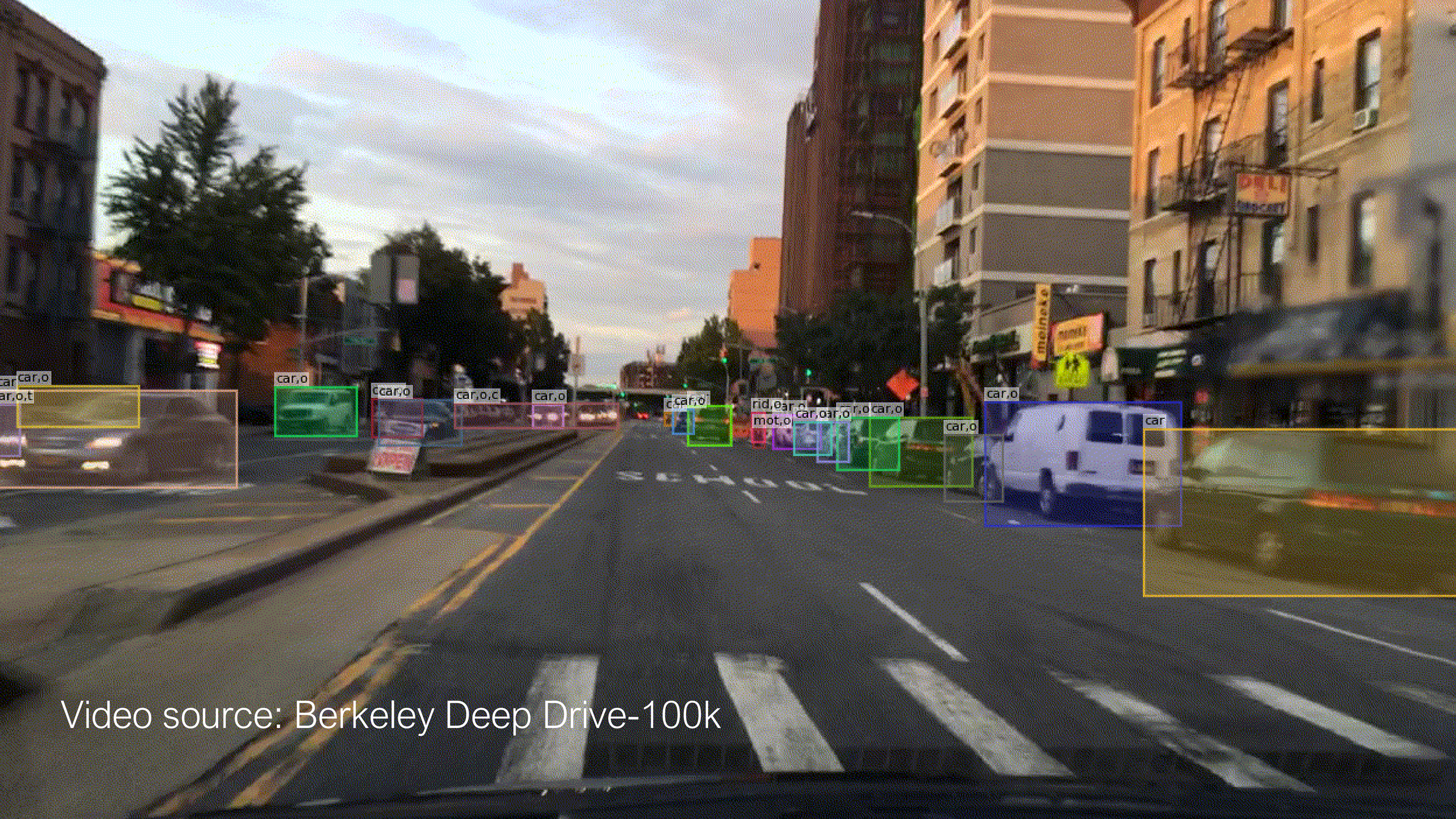
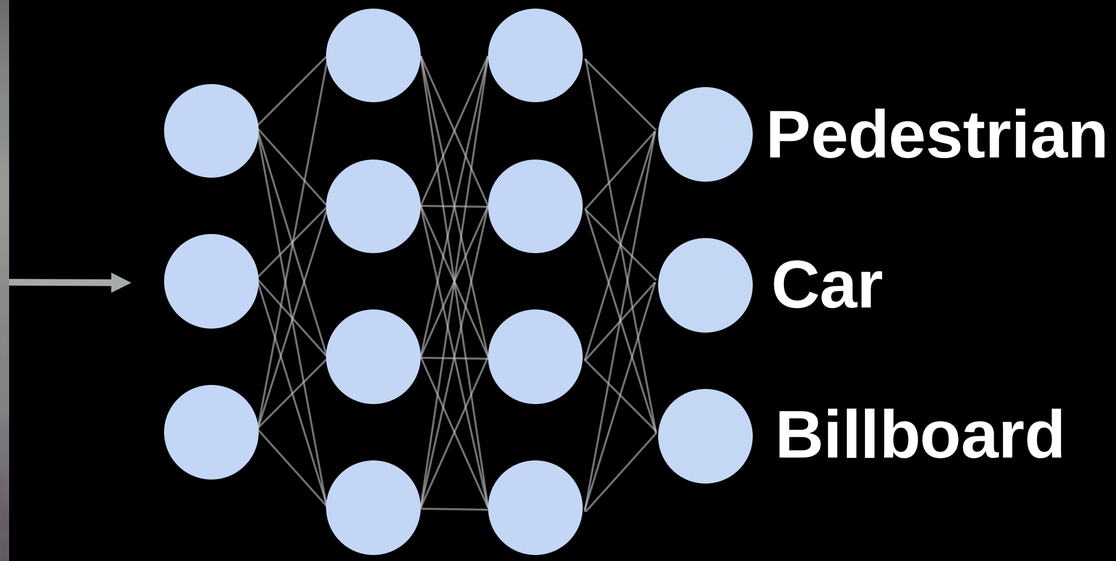
Foundations of Safety-Aware Learning in the Wild

Sharon Li (PI), Jerry Zhu (Co-PI)
Department of Computer Sciences
University of Wisconsin-Madison

@NSF Safe AI Workshop, Feb 26, 2025

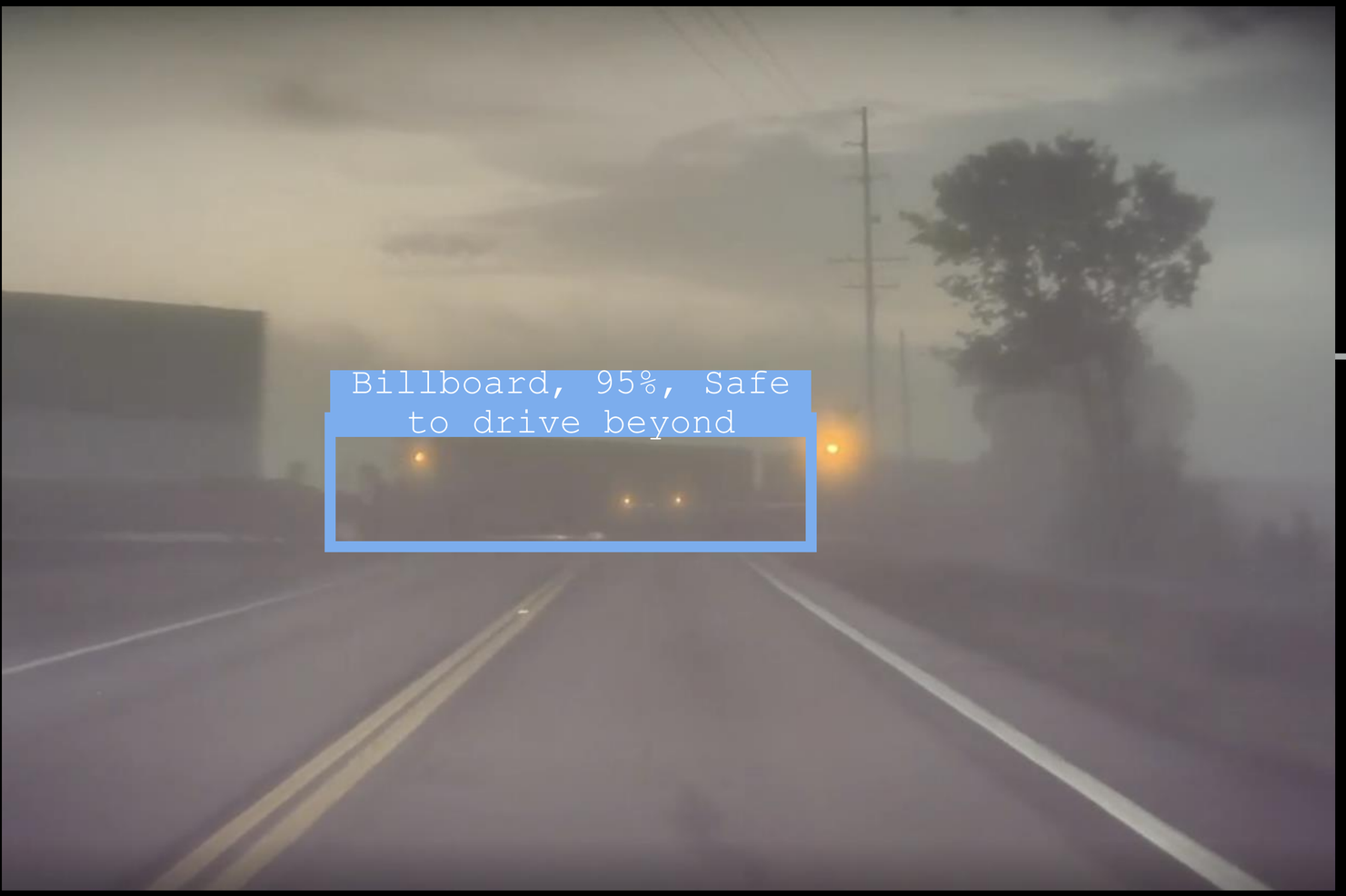

car,o
car,o
car,o
car,o
car,o
car,o
car,o
car,o
car,o
car
Video source: Berkeley Deep Drive-100k

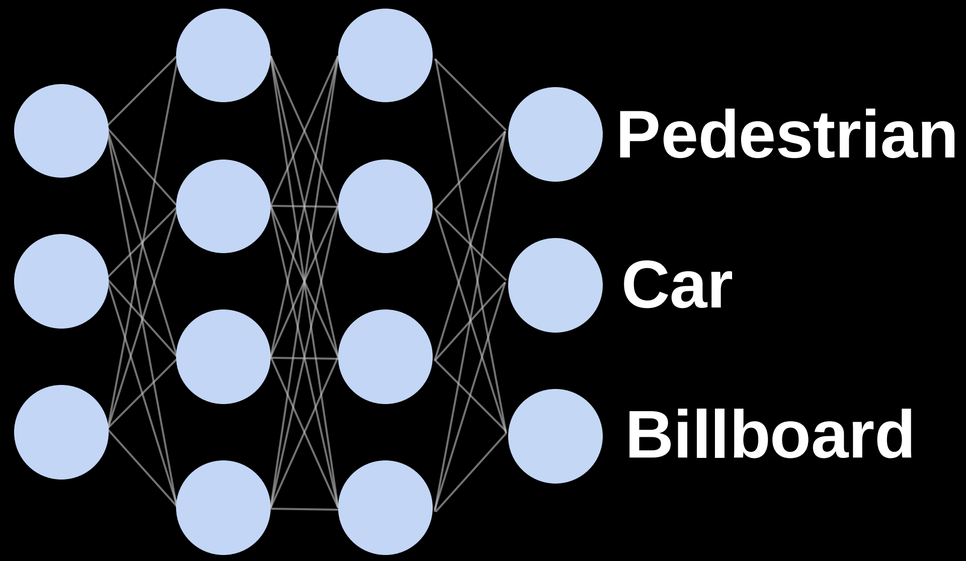
AI models lack awareness of OOD data



https://www.reddit.com/r/Roadcam/comments/1cvq1qp/usa_tesla_vehicle_nearly_runs_into_a_train/

AI models lack awareness of OOD data





Model trained on self-driving dataset produces overconfident predictions for unknown object

yttps://www.reddit.com/r/Roadcam/comments/1cvq1qp/usa_tesla_vehicle_nearly_runs_into_a_train/

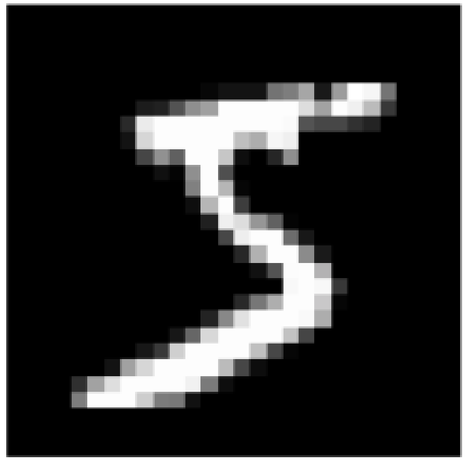
MLLMs Struggle Under Distribution Shifts

What's the woman in the upper left corner of the picture doing?

Multi-modal Large language models

Response

MLLMs Struggle Under Distribution Shifts



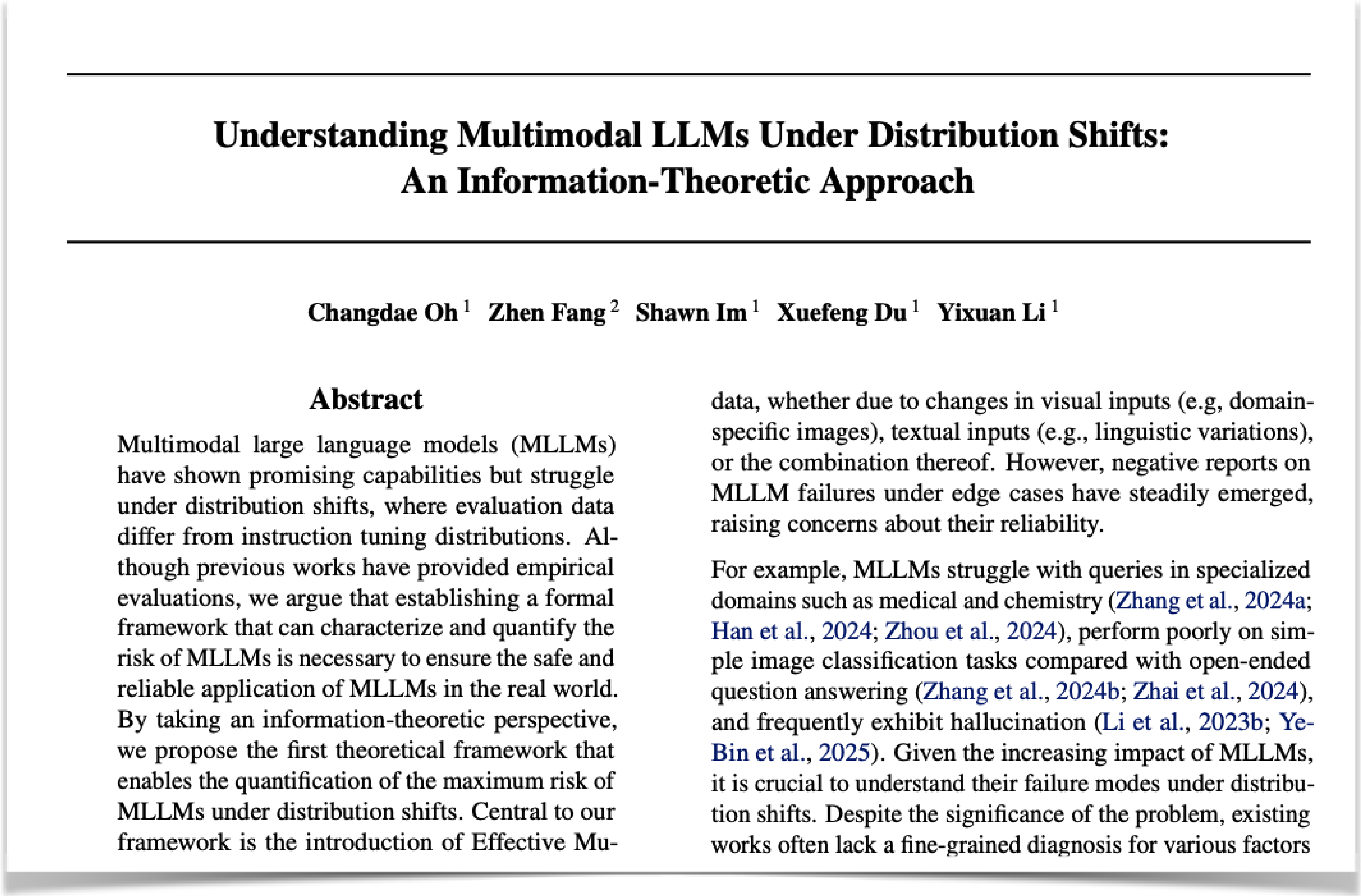
What's the digit in this image?

Multi-modal Large language models

```
The number in the image is 8:
```

Investigating the Catastrophic Forgetting in Multimodal Large Language Models, Zhai et al. 2024

MLLMs Struggle Under Distribution Shifts



Understanding Multimodal LLMs Under Distribution Shifts: An Information-Theoretic Approach

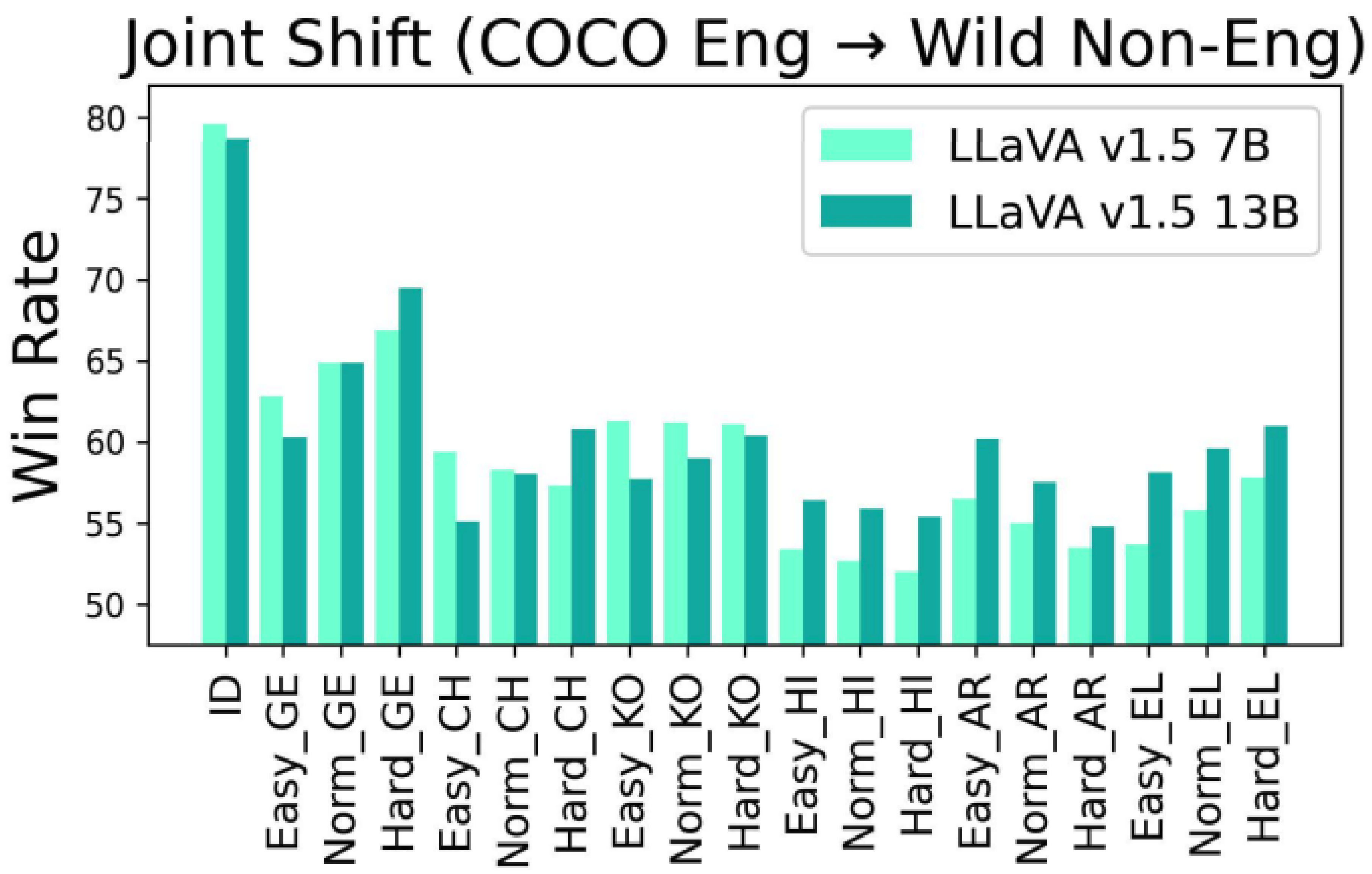
Changdae Oh [1] Zhen Fang [2] Shawn Im [1] Xuefeng Du [1] Yixuan Li [1]

Abstract

Multimodal large language models (MLLMs) have shown promising capabilities but struggle under distribution shifts, where evaluation data differ from instruction tuning distributions. Although previous works have provided empirical evaluations, we argue that establishing a formal framework that can characterize and quantify the risk of MLLMs is necessary to ensure the safe and reliable application of MLLMs in the real world. By taking an information-theoretic perspective, we propose the first theoretical framework that enables the quantification of the maximum risk of MLLMs under distribution shifts. Central to our framework is the introduction of Effective Mu-

data, whether due to changes in visual inputs (e.g, domain-specific images), textual inputs (e.g., linguistic variations), or the combination thereof. However, negative reports on MLLM failures under edge cases have steadily emerged, raising concerns about their reliability.

For example, MLLMs struggle with queries in specialized domains such as medical and chemistry (Zhang et al., 2024a; Han et al., 2024; Zhou et al., 2024), perform poorly on simple image classification tasks compared with open-ended question answering (Zhang et al., 2024b; Zhai et al., 2024), and frequently exhibit hallucination (Li et al., 2023b; Ye-Bin et al., 2025). Given the increasing impact of MLLMs, it is crucial to understand their failure modes under distribution shifts. Despite the significance of the problem, existing works often lack a fine-grained diagnosis for various factors



Theorem 4.5 (Simplified Scenario)**.** *Given an MLLM P_θ and distributions $P_{\mathbf{X}Y}$, $Q_{\mathbf{X}Y}$ which have consistent conditional distributions over variables $X_v|X_t$, $X_t|X_v$, and $Y|\mathbf{X}$, if there exist some constants δ_P and δ_Q such that*

$$D_{\mathrm{JS}}(P_{Y_\theta}\|P_Y) \leq \delta_P, \quad D_{\mathrm{JS}}(Q_{Y_\theta}\|Q_Y) \leq \delta_Q,$$

and denote $P_{Y_\theta} = \mathbb{E}_{P_{\mathbf{X}}}[P_\theta(\cdot|\mathbf{x})]$ and $Q_{Y_\theta} = \mathbb{E}_{Q_{\mathbf{X}}}[P_\theta(\cdot|\mathbf{x})]$, then EMID$(P_{\mathbf{X}Y}, Q_{\mathbf{X}Y}; \theta)$ is upper bounded by

$$\widehat{H}\big(D_{\mathrm{JS}}^{\frac{1}{2}}(P_{X_v}\|Q_{X_v}) + D_{\mathrm{JS}}^{\frac{1}{2}}(P_{X_t}\|Q_{X_t})\big) + 8\Delta^{\frac{1}{4}}, \quad (10)$$

where $\widehat{H} = \max_{\mathbf{x}\in\mathcal{X}}[H(Q_{Y|\mathbf{X}=\mathbf{x}}) + H(P_\theta(\cdot|\mathbf{x}))]$ and $\Delta = \delta_P + \delta_Q$.

Understanding Multimodal LLMs Under Distribution Shifts: An Information-Theoretic Approach. C. Oh, Z. Fang, S. Im, X. Du, Y. Li arXiv 2025.

Our goal: investigate fundamental capabilities and develop new learning **algorithms** and **theories** for *safety-aware learning* in the wild.

Safety should be a built-in objective of ML learning process, not an afterthought or add-on

$$f^* = \text{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{in}}} \mathcal{L}(f(\mathbf{x}_i), y_i), \qquad \text{(Classic ERM, safety-unaware)}$$

Safety should be a built-in objective of ML learning process, not an afterthought or add-on

$$f^* = \text{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{in}}} \mathcal{L}(f(\mathbf{x}_i), y_i), \qquad \text{(Classic ERM, safety-unaware)}$$

$$f^*, g^* = \text{argmin}_{\substack{f \in \mathcal{F} \\ g \in \mathcal{G}}} \frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{in}}} \mathcal{L}(f(\mathbf{x}_i), y_i) + \underbrace{R_{\text{ood}}(g)}_{\text{Risk that discriminates ID vs. OOD}} \qquad \text{(Safety-aware learning)}$$

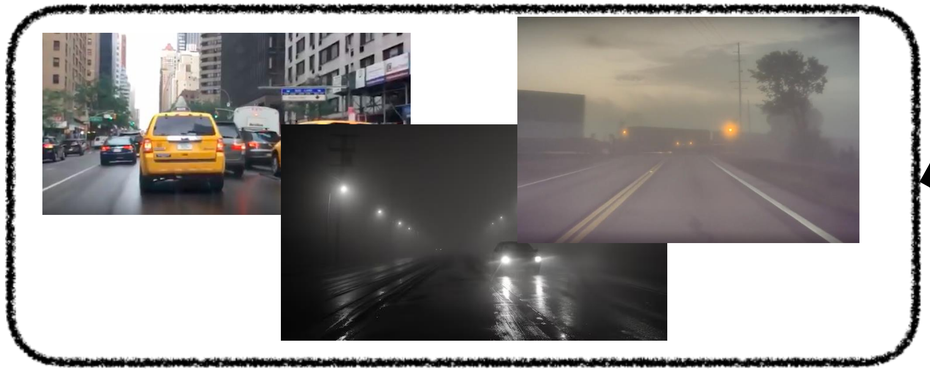
Research Thrusts

T1: Safety-aware learning from offline wild data

T2: Safety-aware learning from online wild data

T3: Safety-aware learning for foundation models

Training environment



Deployment environment

$$f^*, g^* = \text{argmin}_{\substack{f \in \mathcal{F} \\ g \in \mathcal{G}}} \frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{in}}} \mathcal{L}(f(\mathbf{x}_i), y_i) + \underbrace{R_{\text{ood}}(g)}_{\text{Risk that discriminates ID vs. OOD}}$$

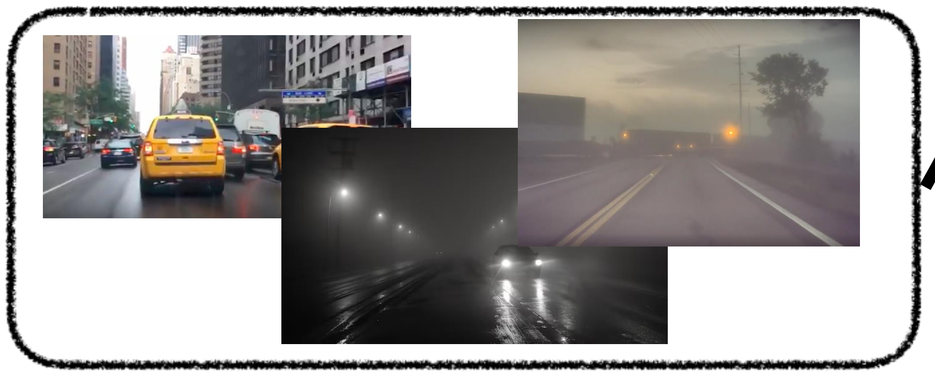
Research Thrusts

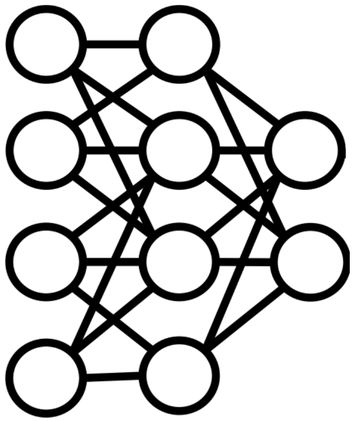
T1: Safety-aware learning from offline wild data

T2: Safety-aware learning from online wild data

T3: Safety-aware learning for foundation models

Training environment





Deployment environment

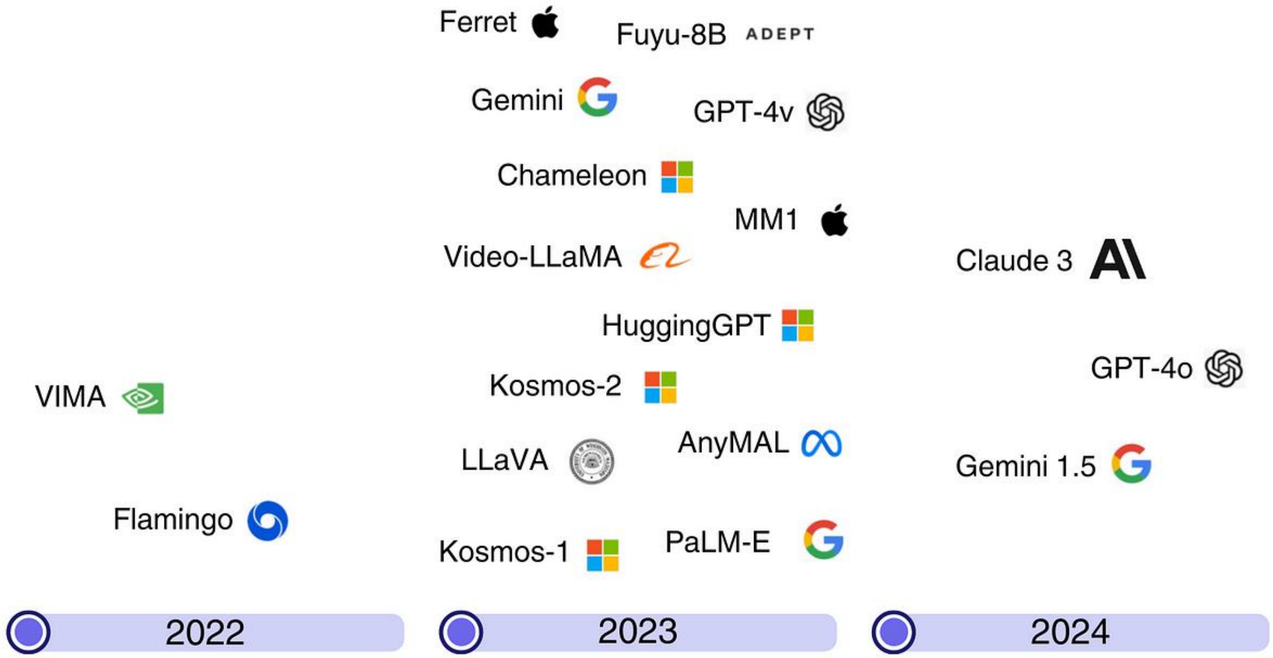
Models’s safety boundary can adapt to dynamic deployment environments!

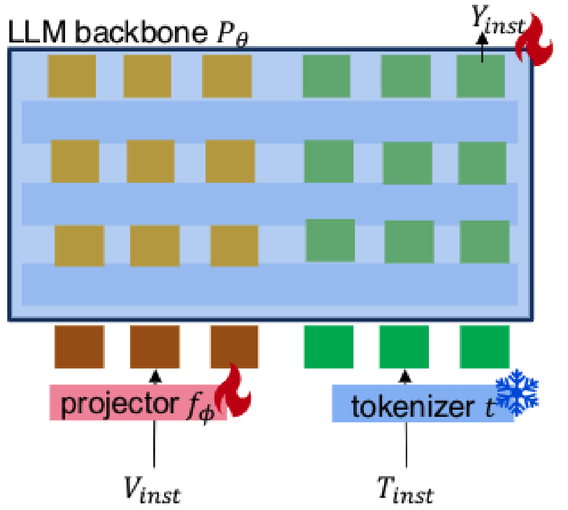
Research Thrusts

T1: Safety-aware learning from offline wild data

T2: Safety-aware learning from online wild data

T3: Safety-aware learning for foundation models





Thank you!